JOHANN WOLFGANG GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

Arbeiten aus dem Institut für Psychologie

der
Johann Wolfgang Goethe-Universität
Frankfurt am Main

Heft 2, 2005

Selektion von Studienbewerbern durch die Hochschulen

Riezlern-Reader XIV

Helfried Moosbrugger, Dirk Frank und Wolfgang Rauch (Hrsg.)

Selektion von Studienbewerbern durch die Hochschulen

Riezlern-Reader XIV

Herausgegeben von
Helfried Moosbrugger, Dirk Frank und
Wolfgang Rauch

Arbeiten aus dem Institut für Psychologie der Johann Wolfgang Goethe-Universität, Heft 2, 2005

Im Rahmen der Frühjahrsuniversität in Riezlern wurden bisher folgende Themen behandelt:

I	(1992)	Methoden der qualitativen Datenanalyse (Erschienen als Riezlern-Reader I, Arbeiten aus dem Institut für Psychologie, Heft 8, 1993)
II	(1993)	Psychologische Evaluationsforschung (Erschienen als Riezlern-Reader II, Arbeiten aus dem Institut für Psychologie, Heft 2, 1994)
III	(1994)	Assessment-Center als Instrument der Personalauswahl und -entwicklung (Erschienen als Riezlern-Reader III, Arbeiten aus dem Institut für Psychologie, Heft 5, 1994)
IV	(1995)	Computergestützte Diagnostik
V	(1996)	Möglichkeiten und Grenzen der wissenschaftlichen Evaluation universitärer Lehre (Erschienen als Riezlern-Reader V, Arbeiten aus dem Institut für Psychologie, Heft 6, 1997)
VI	(1997)	Intervention und Evaluation von Effekten (Erschienen als Riezlern-Reader VI, Arbeiten aus dem Institut für Psychologie, Heft 3, 2002)
VII	(1998)	Methoden der Markt- und Meinungsforschung (Erschienen als Riezlern-Reader VII, Arbeiten aus dem Institut für Psychologie, Heft 1, 1999)
VIII	(1999)	Psychologische Forschung im Internet Möglichkeiten und Grenzen (Erschienen als Riezlern-Reader VIII, Arbeiten aus dem Institut für Psychologie, Heft 10, 1999)
IX	(2000)	Computerbasiertes Experimentieren und Diagnostizieren (Erschienen als Riezlern-Reader IX, Arbeiten aus dem Institut für Psychologie, Heft 4, 2000)
X	(2001)	Psychologische Skalierungsverfahren – Theorie und Anwendungen (Erschienen als Riezlern-Reader X, Arbeiten aus dem Institut für Psychologie, Heft 1, 2001)
XI	(2002)	Visualisierung und Präsentation empirischer Daten (Erschienen als Riezlern-Reader XI, Arbeiten aus dem Institut für Psychologie, Heft 2, 2002)
XII	(2003)	Methoden der Veränderungsmessung (erscheint als Riezlern-Reader XII, in Vorbereitung)
XIII	(2004)	Qualitätssicherung im Bildungswesen (Erschienen als Riezlern-Reader XIII, Arbeiten aus dem Institut für Psychologie, Heft 3, 2004)

Vorwort

„Riezlern XIV“ bestätigt, dass sich unsere Riezlern-Seminare auch im vierzehnten Jahr ihres Bestehens weiterhin großer Beliebtheit erfreuen und eine feste Einrichtung in unserem Ausbildungsprogramm darstellen. Es handelt sich dabei eine Seminarreihe besonderen Stils, welche ich 1992 gemeinsam mit Prof. Dirk Frank, meinem damaligen Mitarbeiter, ins Leben gerufen hatte. Ausschlaggebend waren und sind

- unsere fortgesetzten Anstrengungen zur Verbesserung der Lehre,
- die Überzeugung, dass es im Zeitalter der Massenuniversität neben der Wissensvermittlung auch die Entwicklung der persönlichen Beziehungen zwischen den Studierenden zu pflegen gilt,
- die Existenz unseres universitätseigenen Sport- und Tagungsheimes "Haus Bergkranz" in Riezlern, Kleinwalsertal/Österreich, welches nicht nur über geeignete Seminarräume verfügt, sondern auch noch im Spätwinter beste Gelegenheit zu Wintersport als körperlichem Ausgleich für geistige Betätigung bietet,
- meine Erfahrungen als Dozent auf Sommeruniversitäten der Studienstiftung des Deutschen Volkes, welche mir gezeigt hatten, dass längere Aufenthalte vor Ort eine besonders gute Gelegenheit zur vertieften Befassung mit speziellen Themen bieten, sowie
- die uneingeschränkt positiven Eindrücke aus nun schon dreizehn selbst veranstalteten Riezlern-Seminaren, welche neben dem wissenschaftlichen Ertrag auch deutlich machten, wie sehr Studierende Freiräume zu schätzen wissen, welche sie gemeinsam mit ihren Dozenten verbringen können.

Diese günstige Konstellation von Umständen führte dazu, dass – wiederum in der vorlesungsfreien Zeit – im Lehrangebot des Diplom-Studiengangs Psychologie vom 13. bis 19. März 2005 in Riezlern ein einwöchiges Blockseminar stattfand, in dessen Rahmen wir uns diesmal intensiv mit der Selektion von Studienbewerbern durch die Hochschulen beschäftigen konnten.

Als Ergebnis der seminaristischen Arbeit legen wir nun einen weiteren Band unserer "Riezlern-Reader" vor. Er befasst sich neben einem Einführungskapitel über rechtliche Grundlagen und aktuelle Studien in 17 von den Seminaristen aufbereiteten Kapiteln mit vier großen Themenbereichen, nämlich mit

- methodischen Grundlagen
- Operationalisierungen des Studienerfolgs
- Möglichen Prädiktoren des Studienerfolgs sowie mit
- Beispielen für Studierendenauswahltests.

Meinen Mitherausgebern, Herrn Prof. Dirk Frank und Herrn Dipl.-Psych. Wolfgang Rauch, möchte ich für die vielfältigen Hilfestellungen bei der Literaturauswahl, der Themenstrukturierung und der Redaktion ebenso danken wie für das angenehme Gruppenklima vor Ort in Riezlern, zu welchem sie bei offiziellen wie auch bei inoffiziellen und sportlichen Anlässen nachhaltig beigetragen haben. Herrn Dipl.-Psych. Wolfgang Rauch ist darüber hinaus für die perfekte technische Betreuung im Seminarraum sowie für die wertvolle Hilfe bei der Zusammenstellung und Formatierung der einzelnen Kapitel zum vorliegenden Gesamtwerk zu danken.

Die Beteiligung am diesjährigen Seminar sprengte alle bisherigen Rekorde, waren es doch 21 Teilnehmer aus sechs Nationen, die sich mit dem Seminarthema vertieft befassen wollten. Dabei gebührt unseren Studierenden ein großes Kompliment: Um am Seminar teilnehmen zu können, scheuten sie nicht davor zurück, trotz der vielen Belastungen zu Semesterende auf freiwilliger Basis ihr Thema schriftlich vorzubereiten, für Riezlern eine PowerPoint-gestützte Präsentation anzufertigen, nach dem Vortrag konstruktive Kritik einzuarbeiten und eine schriftliche Endfassung ihrer Beiträge einzureichen. Darüber hinaus erforderte eine ganze Woche des Beisammenseins nicht nur die traditionellen akademischen Höflichkeiten von 17-22 Uhr im Seminarraum, sondern von allen Beteiligten auch ein geneigtes Wohlwollen im übrigen Umgang miteinander, von 8 Uhr beim "obligaten" Frühstück über die vielen Stunden seminaristischer Tätigkeit und gemeinsamen Skilaufens bis hin zu den sehr beliebten nächtlichen Tischtennis-, Spiel- und Hausbarrunden.

Von den Wetterverhältnissen her war unser diesjähriges Blockseminar sehr begünstigt. Alle Tage verwöhnte uns herrlicher Sonnenschein mit angenehmen Temperaturen, so dass wir täglich um 9 Uhr voll Ambition ins Gelände ausrücken und die wunderbaren Tage bei prächtiger Schneelage gründlich auskosten konnten, wobei der Seminarleiter nach seiner langwierigen Unfallverletzung wieder selbst an der Verbesserung und Verfeinerung des schiläuferischen Stils der Seminargruppe unmittelbar und genussvoll mitwirken konnte. Große Verdienste bei der schiläuferischen Betreuung erwarb sich auch Dr. Siegbert Reiß, dem es hierfür vielmals zu danken gilt. Sehr gute Fortschritte machten unsere Anfänger Dorothea M., Heimo D., Samantha W., Sarah S. und Stephanie T., die nach nur wenigen Tagen in der Ski- bzw. Snowboard-Schule z. T. schon tüchtig mit der Gruppe mithalten konnten. Zweifellos waren Christine B. und Nina R. die Cracks. Gruppendynamisch besonders einprägsam waren die nächtliche Geburtstagsparty für Tino K., bei der Beine und Hüften für den nächsten Skifahrtag so richtig locker gemacht wurden, sowie die organisatorische Meisterleistung von Micha D. und Eva R., die uns am Schlussabend doch noch vor dem sicheren Verdursten retteten. Seminaristisch bereichert, sportlich zufrieden, wettermäßig verwöhnt und trotz der großen Teilnehmerzahl ohne jede Verletzung konnten wir schließlich wohlbehalten aus den Alpen zurückkehren.

Insgesamt wurden das Riezlern-Seminar und die Reader-Publikation wiederum mit großer Begeisterung vorbereitet und durchgeführt. Eine Fortsetzung der Frühjahrsuniversität wird von den Studierenden explizit gewünscht, weshalb auch für 2006 geplant ist, weiter zu "readzlern".

Helfried Moosbrugger

Frankfurt am Main, im Juni 2005

Inhaltsverzeichnis

Einführung

Methodische Grundlagen

Operationalisierungen des Studienerfolgs

Mögliche Prädiktoren des Studienerfolgs

Beispiele für Studierendenauswahltests

Studierendenauswahl durch die Hochschulen - rechtliche Grundlagen, empirische Studien und aktueller Stand[1]

Helfried Moosbrugger und Ewa Jonkisz

Hochschulzulassung in Deutschland

Bisherige Gesetzesverordnung

Die Zuständigkeiten für den Hochschulbereich sind in Deutschland zwischen Bund und Ländern verfassungsmäßig geteilt. Der Bund hat die Kompetenz für die Festlegung der allgemeinen Prinzipien für die Gestaltung des Hochschulwesens. Diese Prinzipien sind im Hochschulrahmengesetz des Bundes (HRG) niederlegt. Die Länder hingegen haben die Verantwortung für den laufenden Betrieb der Hochschulen.

Grundlage des aktuellen Hochschulzulassungsrechts ist nach wie vor der Artikel 12 des Grundgesetzes, der allen Deutschen das Recht auf freie Wahl des Berufes, Arbeitsplatzes und der Ausbildungsstätte gewährleistet:

Alle Deutschen haben das Recht, Beruf, Arbeitsplatz und Ausbildungsstätte frei zu wählen. Die Berufsausübung kann durch Gesetz oder auf Grund eines Gesetzes geregelt werden. (Bundestag, o. J., Grundgesetz, Artikel 12)

Durch ein Urteil des Bundesverfassungsgerichts vom 18. Juli 1972, wurde das Recht auf Zulassung zum Hochschulstudium genauer konkretisiert. Das Urteil zielte auf den weiteren Ausbau und die soziale Öffnung der Hochschulen. Es macht deutlich, dass die Regelung der Hochschulzulassung zu den Aufgaben des Staates zählt und dass auch die Berufswahl zu den Freiheitsrechten eines jeden Bürgers gehört. In diesem Urteil traf das Bundesverfassungsgericht die erste Numerus clausus-Regelung, indem dieses Recht als durch Gesetz einschränkbar definiert wurde.

Ein "Numerus clausus" wird unter zwei Bedingungen als gerechtfertigt erachtet, und zwar dann, wenn die Kapazitäten der Bildungseinrichtungen nachweislich erschöpft sind und

---

[1] Bereits erschienen als H. Moosbrugger & E. Jonkisz (2005). *Studierendenauswahl durch die Hochschulen - rechtliche Grundlagen, empirische Studien und aktueller Stand* (Arbeiten aus dem Institut für Psychologie der J. W. Goethe-Universität, Heft 1/2005). Frankfurt am Main: Institut für Psychologie der J. W. Goethe-Universität.

wenn gewährleistet ist, dass die Bewerber nach „sachgerechten“ Kriterien ausgewählt werden. Nach dem Bundesverfassungsgericht entsprechen die Kriterien "Durchschnittsnote" und "Wartezeit" diesen Anforderungen. Da bereits Anfang der 70er Jahre die Anzahl der vorhandenen Studienplätze in manchen Fächern nicht ausreichte, um alle Bewerber zulassen zu können, wurde zum Wintersemester 1973/74 die Zentralstelle für die Vergabe von Studienplätzen (ZVS) eingerichtet. Die ZVS ist eine staatliche Institution, die mit der Auswahl von Studienbewerbern beauftragt wurde. Die ZVS nimmt die Entscheidung nach individuellen Präferenzen des Bewerbers, seiner Leistung und nach sozialen Kriterien (Wohnortnähe, Behinderung, Existenz von Kindern, Ehepartner etc.) vor.

Das Zulassungsverfahren durch die ZVS wurde schon seit mehreren Jahren vielerorts kritisiert, weil häufig nicht die geeigneten Individuen an die Hochschulen kamen. Als Begründung wurde genannt, dass Abiturienten in ihrer schulischen Vorbildung zu heterogen seien (Tent, 1998), dass sich diese Heterogenität in den Abiturnoten nur unzureichend niederschlage bzw. dass die Vergleichbarkeit der Abiturnoten aufgrund unterschiedlicher Fächerkombinationen erschwert sei. Auch die Kursabwahlmöglichkeit von wesentlichen Fächern wurde als Ursache für eine nachlassende Qualifikation der Studienanfänger genannt, ein Zustand, der deutschlandweit beklagt wird (vgl. Rindermann und Oubaid, 1999). Ein anderer Kritikpunkt an dem Hochschulzulassungsverfahren besteht darin, dass die Hochschulen keine Entscheidungsautonomie hätten. Durch die fehlende Entscheidungsfreiheit entstehe ein Zustand von Verantwortungsdiffusion und Verantwortungslosigkeit (Rindermann und Oubaid, 1999). Diesen Einwänden entgegenkommend wurde 2004 das Hochschulrahmengesetz (HRG) novelliert.

Neue Gesetzesverordnung und ihre Folgen

Die Novellierung der rechtlichen Rahmenbedingungen durch das Siebte Gesetz zur Änderung des Hochschulrahmengesetzes (7. HRGÄndG) vom 28. August 2004 bietet den Hochschulen die Möglichkeit, ab dem Wintersemester 2005/2006 60% der zur Verfügung stehenden Studienplätze in den NC-Fächern autonom zu vergeben. Von den übrigen 40% wird jeweils die Hälfte der Studienplätze nach dem Kriterium "Durchschnittsnote der Hochschulzugangsberechtigung" an die Abiturbesten und die andere Hälfte nach dem Kriterium "Wartezeit" vergeben. Dadurch ergibt sich für die einzelnen Länder und Hochschulen zum ersten Mal die Notwendigkeit, Kriterien für die Vergabe der Studienplätze in den Fächern mit einer Zulassungsbeschränkung zu benennen.

Das 7. HRGÄndG beschreibt einen Katalog von sechs spezifischen Kriterien, die zur Studierendenauswahl herangezogen werden dürfen. Zu den sechs im HRG genannten Kriterien zählen (Bundesministerium für Bildung und Forschung, 2004):

a) Grad der Qualifikation (Durchschnittsnote der Hochschulzugangsberechtigung),

b) gewichtete Einzelnoten (fachspezifische Eignung),

c) Ergebnis eines fachspezifischen Studierfähigkeitstests,

d) Art der Berufsausbildung oder Berufstätigkeit,

e) Ergebnis eines von der Hochschule durchzuführenden Gesprächs und

f) eine Verbindung der Kriterien a bis e.

Auf der Basis des HRG sollten die 16 Länder ihre Hochschulgesetze, in denen weitere Einzelheiten geregelt werden und ggf. dieser Kriterienkatalog modifiziert wird, bis Ende 2004 erlassen. Die Umsetzung der entsprechenden Änderungen in den landesrechtlichen Vergabeordnungen lässt aber noch auf sich warten.

In den Satzungen der einzelnen Hochschulen wird das jeweilige Zulassungsverfahren im Einzelnen geregelt. In der Satzung der Johann Wolfgang Goethe-Universität sind bspw. die Kriterien und die Entscheidungen für die Auswahl von Studienbewerberinnen und Studienbewerbern in zulassungsbeschränkten Studiengängen festgehalten. Die Auswahlkriterien entsprechen hier denen, die im 7. HRGÄndG als mögliche Optionen aufgeführt sind. Im Anhang sind die Regelungen für die einzelnen Studiengänge beschrieben. Bei dem Zulassungsverfahren für den Diplomstudiengang Psychologie bspw. wird gegenwärtig auf Informationen aus dem Abiturzeugnis zurückgegriffen, im Fach Pharmazie ist der Einsatz von Auswahlgesprächen vorgesehen.

Chancen für die Hochschulen

Die den Hochschulen eingeräumte Möglichkeit zur autonomen Auswahl der Studierenden eröffnet die Chance, Hand in Hand mit einer größeren Profilbildung und Diversifizierung der Studienangebote, die Selektion der dazu passenden Bewerber selbst vorzunehmen. Einerseits können die Hochschulen maßgeschneiderte Studienangebote machen und andererseits können sich die Studierende Studiengänge aussuchen, die auf ihr spezielles Fähigkeitsprofil zugeschnitten sind. Durch ein gezieltes Beratungsangebot von Seiten der Universitäten in Form von Realistic Job Previews und durch Selbstselektion (s. u.) können die Studierenden bei der Wahl des für sie passenden Studiengangs unterstützt werden, indem ihre Eignung bzw. Nichteignung für den speziellen Studiengang schon vor der Aufnahme des Studiums prognostiziert wird. In Anlehnung an den Person-Job-Fit-Ansatz von Amelang (1997) wird auf diese Weise von vornherein die größtmögliche Passung zwischen Studienangebot und Studierenden erreicht und die Gefahr von Fehlentscheidungen reduziert.

Insgesamt lassen sich fünf grundsätzliche Ziele unterscheiden, welche mit dem Einsatz von Auswahlverfahren verfolgt werden sollen (vgl. Arnhold & Hachmeister, 2004):

Profilbildung

Zum einen handelt es sich um die Profilbildung der Hochschulen, der Fachbereiche und der einzelnen Studiengänge. Die Hochschule bzw. der Fachbereich entwickelt ein Anforderungsprofil, für das die passenden Studierenden zu identifizieren sind. Voraussetzung dafür ist die Klärung der Frage, wie das jeweilige Profil auszusehen hat und welche Fähigkeiten und Fertigkeiten die angehenden Studierenden mitbringen müssen. Das Profil eines Studiengangs lässt sich anhand drei zentraler Dimensionen beschreiben.

- Das Leistungsanforderungsprofil, nämlich die Forderung der Hochschulen, sich die "besten" Studierenden aussuchen zu dürfen, steht gegenwärtig im Vordergrund. Die Qualifikation der Studierenden wird an solchen Kriterien wie kognitive Leistungsfähigkeit, Leistungsmotivation oder Leistungsstärke in Schlüsselqualifikationen (Teamfähigkeit, Kommunikationsfähigkeit u. ä.) festgemacht.
- Das inhaltliche Profil eines Studiengangs ist durch unterschiedliche inhaltliche Ausrichtungen und Schwerpunktbildungen innerhalb der Fächer gekennzeichnet, z. B. durch eine Spezialisierung auf Werbe- oder Forensische Psychologie.
- Das strukturelle Profil schließlich wird an solchen Größen wie Abschlussart (Diplom, Magister, Bachelor, Master), Studienorganisation (Präsenz-, Fern-, Vollzeit- und Teilzeitstudium), obligatorischen Praktika und / oder Auslandssemestern festgemacht.

Homogenisierung des Leistungs- bzw. Vorkenntnisniveaus

Mit der Profilbildung hängt die Homogenisierung des Leistungs- bzw. Vorkenntnisniveaus innerhalb des Studiengangs eng zusammen. Diese bietet der Hochschule die Möglichkeit, ihr Angebot optimal auf die Studierenden auszurichten. Das Ziel besteht demnach nicht darin, die durchschnittlichen Leistungs- und Vorkenntnisniveaus der Bewerber unter Beibehalt einer beträchtlichen Variation zu erhöhen, sondern ein einheitlicheres variationsarmes Leistungs- und Vorkenntnisniveau auf einer angemessenen Höhe anzustreben. So ist eine Spezialisierung von Studiengängen auf eine bestimmte Klientel vorstellbar.

Senkung der Abbrecherquote

Durch die gegenseitige Selektion, sowohl seitens der Bewerber als auch seitens der Hochschule, kann ein höheres Maß an gegenseitiger Wertschätzung bzw. Bindung erzielt werden. Diese gegenseitige Bindung ist umso höher, je selektiver das Auswahlverfahren gestaltet worden ist. Auf der Seite der Studierenden führt dies zu einer Senkung der Ausfallquote durch Studienabbruch oder Hochschulwechsel und zu einem höheren Engagement für das Studium und die Hochschule insgesamt. Auch auf der Seite der

Lehrenden ist ein höheres Engagement für die nach eigenen Maßstäben ausgewählten Studierenden zu erwarten.

Studiumszulassung

Bei dem Wettbewerb um besonders fähige Studierende darf nicht aus den Augen verloren werden, dass das allgemeine Ziel eines Auswahlverfahrens in der Zulassung zum Studium besteht und nicht in einer Ausgrenzung vom Studium. Ein zu aufwendig gestaltetes Verfahren könnte u. U. erschweren, eine angemessne Anzahl von passenden Studierenden zu finden. Die Entwicklung des Anforderungsprofils des Studiengangs sollte deshalb interaktiv verlaufen, damit eine ausreichende Anzahl von Bewerbern gewährleistet werden kann. Eine Hochschule muss sowohl die eigenen Angebote und ihre Qualität, als auch die Population der potentiellen Studierenden im Blick behalten und ggf. die Auswahlziele und die Kriterien anpassen.

Berufsqualifizierende Abschlüsse

Schließlich bietet die spezifische Auswahl der Studierenden für bestimmte Studiengänge den Hochschulen die Möglichkeit, sich vom Konzept der allgemeinen Studierfähigkeit zu verabschieden und die Eignung für konkrete berufsqualifizierenden Abschlüsse wie BA/BSc festzustellen. Damit wird ein weiterer Schritt in Richtung einer in der Öffentlichkeit geforderten “Modernisierung” des Hochschulwesens und Verkürzung der Studiendauer getan.

Eignungsbeurteilung im Kontext der Auswahl von Studierenden

Mögliche Prädiktoren des Studienerfolgs

Für die Realisierung der im vorangegangenen Abschnitt beschriebenen Ziele der Studierendenauswahl ist eine angemessene Prädiktorenauswahl von entscheidender Bedeutung. Die dahinter liegende Grundannahme geht von der Existenz personenbezogener, stabiler, messbarer und prognosefähiger Merkmale aus, welche in Zusammenhang mit Studienerfolg stehen. Für die Erfassung dieser Merkmale steht eine Reihe von Verfahren zur Verfügung.

Schulnoten

Schulnoten stellen eine leicht verfügbare Informationsquelle dar, da sie zumindest bei inländischen Bewerbern immer vorhanden sind. Sie geben Hinweise über den Leistungs- bzw. Wissensstand der Bewerber, aber auch über seine kognitiven Fähigkeiten. Darüber hinaus spiegeln die Schulnoten solche Faktoren wie Leistungsmotivation, Fleiß, Anpassung, Wissensmanagement und Konzentrationsfähigkeit wieder. Die

Abiturdurchschnittsnote ist durch das höhere Aggregationsniveau messgenauer als die Einzelfachnoten. Auch die prognostische Validität des Gesamtdurchschnitts der Schulabschlussnoten ist höher als die der Einzelfachnoten und liegt zwischen .28 und .48 (Rindermann & Oubaid, 1999, S. 178); durch die zusätzliche Berücksichtigung von Einzelfachnoten können jedoch inkrementelle Validitäten erzielt werden.

Für den Einsatz der Schulnoten spricht insgesamt nicht nur ihre hohe prognostische Validität für den Studienerfolg, sondern auch die leichte Zugänglichkeit, die schwierige Verfälschbarkeit und die juristische Unangreifbarkeit (vgl. DGPs, 2004).

Tests

Zur Erhebung von Kompetenzen und Persönlichkeitsmerkmalen können unterschiedliche Arten von Studieneingangstests eingesetzt werden. In Schulleistungstests werden schulbezogene Kenntnisse (z. B. Mathematik, Deutsch) erhoben mit dem Ziel, die Lehrerurteile zu objektivieren und gewisse Mindeststandards zu setzen. Studienfachspezifische Kenntnistests sind auf spezielle Inhalte, wie bspw. Fremdsprachen (Test of English as a Foreign Language; TOEFL; Educational Testing Service, 2005), begrenzt. Studierfähigkeitstests messen solche Fähigkeiten, die für eine Bewältigung der Studienanforderungen notwendig sind, jedoch nicht durch Schulleistungen oder Schulleistungstest erfasst werden. Generell lassen sich hier allgemeine und studienfachspezifische Tests unterscheiden. Beispiele für die erste Variante sind der in den USA eingesetzte Scholastic Aptitude Test (SAT; Educational Testing Service, 2005) und der Graduate Record Examination (GRE; Educational Testing Service, 2005). Der in Deutschland entwickelte, aber mittlerweile nicht mehr durchgeführte Test für Medizinische Studiengänge (TMS; Institut für Test- und Begabungsforschung, 1986) gehört zu den bekanntesten fachspezifischen Fähigkeitstests. In Persönlichkeitstests geben die Bewerber Selbsteinschätzungen hinsichtlich eigener Motivation, Interessen, Arbeitshaltungen, Verhaltensweisen etc. ab. Im Gegensatz zu Leistungstests besteht hier aber die Gefahr einer absichtlichen Verfälschung der Angaben zugunsten der sozialen Desirabilität.

Interviews

Mit Auswahlinterviews können solche Faktoren wie kommunikative Kompetenzen, individuelle Besonderheiten und Studienmotivation erfasst werden. Interviews sind jedoch meist sehr subjektiv, anfällig für Urteilsfehler und in hohem Maße verfälschbar. Generell ist ihre Objektivität, Reliabilität und somit auch ihre prognostische Validität wesentlich geringer als gemeinhin angenommen. Darüber hinaus sind sie mit einem erheblichen personellen und zeitlichen Aufwand verbunden. Befürchtet wird zudem, dass "durch Auswahlgespräche die bereits sehr hohe soziale Selektivität von Studierenden in Deutschland verstärkt wird" (DGPs, 2005, S. 153).

Bewerbungsschreiben, Aufsätze, Essays

Auch mit diesen Verfahren lassen sich Studieninteressen und Studienmotivation erfassen. Sie können aber auch Auskunft über Ausdrucksfähigkeit und Sprachbeherrschung, Berufserfahrung und allgemeinen Bildungshintergrund des Bewerbers geben. Noch mehr als bei einem Auswahlinterview besteht hier die Gefahr zugunsten der sozialen Erwünschtheit ein geschöntes Bild von sich selbst zu entwerfen.

Probezeit

Analog zu den Arbeitsproben in der Berufswelt liegt der Probezeit für Studierende die Annahme zugrunde, dass die zukünftige Leistung bzw. ein erfolgreicher Studienabschluss sich am besten aufgrund der Leistungen in der Studieneingangsphase vorhersagen lässt. Obwohl sich dies empirisch zeigen lässt (Vorhersage des Studienerfolgs anhand der Vordiplomsnote, s. Reiß und Moosbrugger, 2004), muss hierbei der mit einer Probezeit einhergehende Zeit- und Ressourcenverlust bedacht werden.

Selbstselektion

Eine vorausgehende Selbstselektion vermindert den nachfolgenden Selektionsaufwand durch die Hochschule dadurch, dass sich im Idealfall nur mehr geeignete Kandidaten tatsächlich bewerben. Zum einen lässt sich Selbstselektion durch Information erzielen. Ausgehend von einer bestimmten Zielgruppe kommuniziert die Hochschule, der Fachbereich bzw. der Studiengang das eigene Profil und die Anforderungen, die zu Beginn und während des Studiums an den Studierenden gestellt werden ("Realistic Job Preview"). Um den Bewerbern zu erleichtern, die eigenen Fähigkeiten, Fertigkeiten, Kenntnisse und Persönlichkeitsmerkmale mit den im Anforderungsprofil zu vergleichen, kann ihnen ein Selbst-Test angeboten werden (z. B. webbasiertes SelfAssessment-Tool bei RWTH Aachen; verfügbar unter: http://www.assess.rwth-aachen.de). Zum anderen können die Hochschulen und Fachbereiche sich Selbstselektion durch gesteigerte Anforderungen zunutze machen. Hier werden der Aufwand bzw. die Anforderungen an den Bewerber so hoch gesetzt, dass nur besonders motivierte und befähigte Bewerber überhaupt die Unterlagen zusammenstellen (bspw. Praktika, Essays, Künstlermappen).

Kriterien des Studienerfolgs

Die Feststellung des Studienerfolgs kann an verschiedenen Kriterien festgemacht werden. Bei der Auswahl der Kriterien muss ein besonderes Augenmerk auf deren Qualität und auf das Aggregationsniveau, das vergleichbar mit dem der Prädiktoren sein soll, gelegt werden.

Studienabschluss und Studienabbruch

Ein abgeschlossenes Studium stellt ein erstes, allerdings sehr "grobkörniges" Erfolgskriterium dar. In Umkehrschluss gilt ein Hochschulabgang ohne Erreichung des

vorher angestrebten Studienabschlusses als ein besonders hartes Kriterium des Studienmisserfolgs. Ein Studienabbruch zieht Konsequenzen auf individueller, institutioneller und staatlicher Ebene nach sich (Gold, 1988). Für das Individuum bedeutet er eine persönliche Niederlage und einen Verlust von Zeit und Geld, für die Hochschulen und den Staat einen unnötigen Verbrauch an Ressourcen.

Hauptdiplomsnoten oder Vordiploms-/ Zwischenprüfungsnoten

Nach Baron-Boldt, Schuler und Funke (1988) werden Studienabschlussnoten oder Zwischenprüfungsnoten am häufigsten als Kriterien des Studienerfolgs herangezogen. Sie sind leicht erfassbar, verfügen über fachspezifische Vorhersagekraft für die Arbeitsmarktchancen (Gold & Souvignier, 1997) und stellen damit das wichtigste quantitative Kriterium dar.

Studiendauer

Vor dem Hintergrund der Debatten um Langzeitstudierende und Langzeitstudiengebühren gewinnt die Studiendauer als ein Erfolgskriterium immer mehr an Bedeutung. Ziel ist es, in kurzer Zeit zu einem qualifizierten Abschluss zu gelangen. Die Erfassung dieser Variablen bringt allerdings Schwierigkeiten mit sich, da Quereinsteiger-, Fach-, Ortswechsler und Teilzeitstudierende die Statistiken verzerren.

Studienzufriedenheit

Studienerfolg wird nur selten anhand der Studienzufriedenheit erfasst. Eine der Ursachen liegt zweifellos darin, dass Studienzufriedenheit in vielfältiger, gelegentlich widersprüchlicher Weise operationalisiert werden kann. Wird die Studienzufriedenheit während des Studiums erfragt, ist sie ein Indikator für die aktuelle Befindlichkeit. Wenn sie aber retrospektiv erhoben wird, so wird mit ihr der eigene Studienerfolg bewertet (Gold & Souvignier, 1997).

Allgemeine berufsqualifizierende Kompetenzen

Allgemeine Kompetenzen wie Kommunikationsverhalten, Führungsqualitäten, Zeitmanagement und andere sog. Schlüsselqualifikationen sind in der Berufswelt gewiss von großer Bedeutung; aufgrund ihrer Definitions- und Messproblematik werden sie jedoch zur Feststellung des Studienerfolgs bisher selten verwendet.

Berufserfolg

Berufserfolg sollte das Erfolgskriterium per se darstellen. Seine Messung ist aber mit einer Vielzahl von Problemen behaftet: angefangen von der Definition (Arbeitsmarktchancen, Position, Einkommen, Vorgesetztenurteil, Zufriedenheit) bis zu zeitlichem Abstand zwischen den Messzeitpunkten.

Bedingungsmodell des Studienerfolgs

Als Zusammenschau zwischen Prädiktoren und Kriterien findet sich bei Rindermann und Oubaid (1999) ein Modell der Bedingungen des Studienerfolgs (s. Abbildung 1). Die individuellen Eingangsqualifikationen der Studienanfänger (links) stellen dabei die wichtigste Determinante des Ausbildungs- und Berufserfolgs (rechts), in diesem Sinne, dass ihre positive Ausprägung die Wahrscheinlichkeit eines positiven Outcomes erhöht. Je nachdem welche individuellen Prädiktoren und Kriterien des Studienerfolgs gewählt werden, können die Ergebnisse u. U. durchaus unterschiedlich ausfallen. Zu den Bedingungen des Studienerfolgs gehört auch eine Reihe von institutionellen Variablen, nämlich die allgemeine Studien- und Lehrqualität der jeweiligen Hochschule (oben) sowie die jeweiligen gesellschaftlichen Rahmenbedingungen (unten).

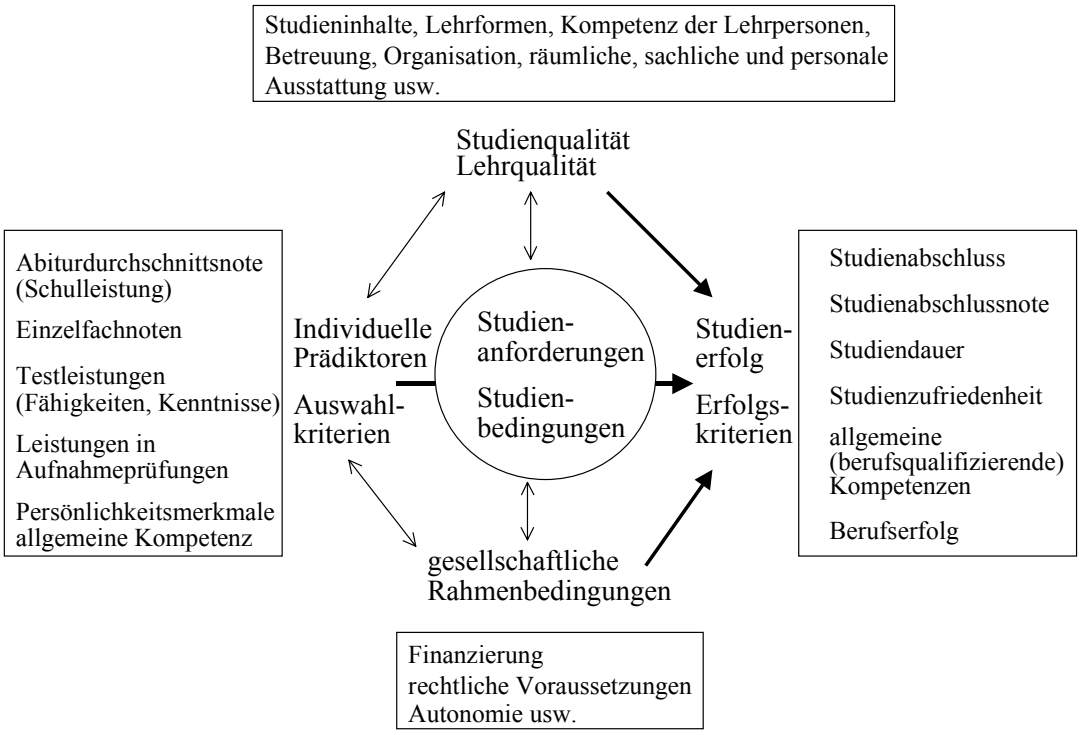


Abbildung 1: Bedingungsmodell des Studienerfolgs (aus Rindermann & Oubaid, 1999, S. 176)

Empirische Studien

Will man empirische Daten über die Bedingungen des Studienerfolgs gewinnen, so können in erster Linie Absolventenstudien herangezogen werden. Darüber hinaus lassen sich aus solchen Untersuchungen auch Kenntnisse über den Verbleib von Absolventen bzw. über deren Übergang in das Berufsleben ableiten. Im Folgenden wird auf zwei solche, kürzlich

durchgeführte Studien für Absolventinnen und Absolventen in Psychologie eingegangen, nämlich auf die Studie der Deutschen Gesellschaft für Psychologie (Schneller & Schneider, 2005), welche mehr an der beruflichen Situation der Absolventinnen und Absolventen interessiert war und auf die Studie der Universität Frankfurt ((Reiß & Moosbrugger, 2004; Moosbrugger, Reiß, & Eisenhuth, 2005), welche in breiter Form auch Eignungs-qualifikationen und Studienbedingungen erfasste.

Studie der Deutschen Gesellschaft für Psychologie (DGPs)

Von der Deutschen Gesellschaft für Psychologie (DGPs) wurde zwischen Februar und Juni 2004 eine bundesweite Befragung der Absolventinnen und Absolventen des Jahres 2003 im Studiengang Psychologie durchgeführt (Schneller & Schneider, 2005). Inhalte dieser Befragung bezogen sich auf allgemeine Angaben zum Studium, Informationen zur Stellensuche und auf die momentane berufliche Situation der Absolventinnen und Absolventen. Die Probanden wurden entweder per Post oder per Email rekrutiert. An der Untersuchung beteiligten sich 1084 Absolventinnen und Absolventen (Rücklaufquote 58%) mit einem Durchschnittsalter von 29,9 Jahren (Standardabweichung 5,61), davon 260 Männer und 824 Frauen. Bei den folgenden Angaben ist die Möglichkeit einer mangelnden Repräsentativität der Stichprobe und damit einhergehende unbekannte Selektionsprozesse nicht auszuschließen.

Angaben zum Psychologiestudium: Im Durchschnitt betrug die Vordiplomsnote der Absolventinnen und Absolventen 2,0 bei einer Standardabweichung von 0,64, die Hauptdiplomsnote lag im Mittel bei 1,54 (Standardabweichung 0,43). Bezüglich der Studienzufriedenheit, waren viele aus retrospektiver Sicht der Meinung, dass das Studium mehr Praxisnähe haben sollte und mehr als die Hälfte wünschten sich ein höheres Engagement der Dozenten. Dennoch standen über drei Viertel dem Psychologiestudium positiv entgegen und über die Hälfte der Teilnehmer waren sich sicher, Psychologie wieder studieren zu wollen.

Angaben zur momentanen beruflichen Situation: 73,8% der Befragten hatten zum Zeitpunkt der Befragung eine feste Anstellung, 12,9% befanden sich definitiv auf der Suche, 3,3% hatten eine zugesicherte Stelle. Die restlichen 10% gingen keiner Erwerbstätigkeit nach, wobei als Ursache oft eine Ausbildung oder Kindererziehung genannt wurde. Laut Angaben ist fast ein Drittel der Personen mit qualifiziertem Arbeitsplatz an einer Universität tätig. Weitere Arbeitgeber sind unter anderem Kliniken, Vereine, Kommunen, private Praxen und Kirche.

Angaben zur Bezahlung und Zufriedenheit mit der Arbeitsstelle: 40% der erwerbstätigen Personen mit qualifizierten Arbeitsplätzen sind mit ihrer Bezahlung wenig bis überhaupt nicht zufrieden, mit den Arbeitsinhalten sind dagegen 80% zufrieden. Das Gefühl, die beruflichen Vorstellungen ziemlich bzw. vollkommen verwirklicht zu haben, hatten mehr

als 50%. Auch in Bezug auf ihre Jobsicherheit, die Perspektive der zukünftigen Bezahlung und die Entfaltungsmöglichkeiten waren die meisten zufrieden.

Frankfurter Absolventenstudie

Im Vergleich zu der DGPs-Absolventenstudie liefert die Frankfurter Absolventenstudie (Reiß & Moosbrugger, 2004; Moosbrugger, Reiß, & Eisenhuth, 2005) mehr Hinweise für Prädiktoren der Studiendauer und des Studienerfolgs. Die Datenerhebung erfolgte im Zuge der Diplomarbeit von Eisenhuth (2004) und wurde mittels des zu diesem Zweck entwickelten Frankfurter Psychologie-Absolventen-Fragebogens durchgeführt (Moosbrugger & Eisenhuth, 2004). Diese retrospektive Untersuchung richtete sich an alle (N = 212) Absolventinnen und Absolventen des Diplom-Studiengangs Psychologie der J. W. Goethe Universität, die vom Wintersemester 1995/96 bis einschließlich Sommersemester 2002 ihr Studium abgeschlossen hatten. Insgesamt waren 142 Personen (Rücklaufquote 67%) im Alter von 25 bis 62 Jahren an der Studie beteiligt, davon 24 Männer und 118 Frauen, von denen ausgefüllte Fragebögen retourniert wurden. Im Durchschnitt betrug die Hauptdiplomsnote der Befragten 1,69, im Vergleich zu 1,76 - der Durchschnittsnote aller Absolventinnen und Absolventen des Studiengangs Psychologie in Frankfurt.

Als Kriterien wurden "Studiendauer" und "Studienerfolg" verwendet. Darüber hinaus wurde die berufliche Situation der Absolventinnen und Absolventen erfasst. Die Korrelation zwischen den beiden Kriterien Studiendauer und Studienerfolg beträgt 0,33 und ist statistisch signifikant. Der FPAF erfasst in detaillierten Fragen Informationen zu „individuellen Voraussetzungen“ wie durchschnittliche Note der Hochschulzugangsberechtigung, EDV-Kenntnisse zu Studienbeginn und Wichtigkeit einer kurzen Studiendauer, „Studienbedingungen“ sowie „Studienergebnisse im Vordiplom“.

Als Prädiktoren des Studienerfolgs, operationalisiert durch die durchschnittliche Hauptdiplomsnote, erwiesen sich vor allem ein niedriges Studienalter, gute Vordiplomsnoten, ein Engagement in Gremien, eine enge Betreuung durch die Mitarbeiter und die Nutzung von Vorlesungen und Selbststudium bei der Vorbereitung auf die Hauptdiplom-Prüfung als günstige Prädiktoren (multiple Regressionsanalyse mit 56% Varianzerklärung).

Was die Studiendauer betrifft, so wird diese gemäß den Ergebnissen dieser Studie durch eine gute Abiturs- und Vordiplomsnote sowie gute EDV- und Internetkenntnisse zu Beginn des Studiums verkürzt. Studienverlängernd wirkt sich vor allem das Vorhandensein von Kindern aus. Des Weiteren spielt die individuelle Einstellung zur Studiendauer eine entscheidende Rolle, denn je wichtiger die Studiendauer in den Augen des Studierenden ist, desto geringer ist diese. Auch eine enge Bindung zum Lehrpersonal wirkt sich günstig auf die Studiendauer aus (multiple Regressionsanalyse mit 36% Varianzerklärung).

Bei der Betrachtung der beruflichen Situation der Absolventinnen und Absolventen zeigte sich, dass der überwiegende Teil von ihnen in den beruflichen Einsatzfeldern "Klinische

Psychologie/Psychotherapie" (27%), "Arbeits- und Organisationspsychologie" (21%) und "Forschung und Lehre" (19%) beschäftigt ist. Nur 10% der Befragten hatten bis zum Erhebungsabschluss keine Stelle gefunden, 20% gaben an, selbständig und 72% nicht selbständig zu sein.

Aktueller Stand

DIN 33430

Für die berufsbezogene Eignungsbeurteilung existiert seit Kurzem die DIN 33430 über "die Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen" (Deutsches Institut für Normierung, 2002). Sie stellt ein Markenzeichen für qualitätsorientierte laufbahnbezogene Entscheidung dar. DIN 33430 "stellt Qualitätsstandards zu Inhalt und Ablauf von Verfahren für die Personalauswahl und die Personalentwicklung und zur Qualifikation der internen oder externen Auftragnehmer auf" (Westhoff, Hellfritsch, Hornke, Kubinger, Lang, Moosbrugger, Püschel & Reimann, 2004, S. 15). Diese Norm soll zu einer "fachgerechten Entwicklung und zum sachgerechten Einsatz von Verfahren der Eignungsbeurteilung" beitragen, sowie einer "kontinuierlichen Verbesserung der Verfahren zur Eignungsbeurteilung" dienen (DIN, 2002). Da die Studierendenauswahl eine Form der Eignungsbeurteilung darstellt, sollten die beschriebenen Qualitätskriterien auch auf diese Situation übertragen werden. Deshalb sollten die am Auswahlverfahren Beteiligten die vom Testkuratorium der Föderation Deutscher Psychologenvereinigungen abgenommene Lizenzprüfung nach DIN 33430 nachweisen können.

Die in der DIN 33430 festgelegten Normen machen deutlich, dass insbesondere dem Fach Psychologie die Kompetenz und Erfahrung zukommt, eine qualitativ hochwertige Eingangsdiagnostik zu entwerfen, durchzuführen und zu evaluieren. Diese Expertise hat sich bereits bei der Entwicklung des Tests für Medizinische Studiengänge (TMS; Institut für Test- und Begabungsforschung, 1986) bewährt, einem fachspezifischen Studienfähigkeitstest, der aufgrund veränderter Bedingungen allerdings mittlerweile wieder eingestellt wurde.

Kommission innerhalb der DGPs

Von dem Vorstand der DGPs wurde im Oktober 2004 eine Kommission eingerichtet und setzt sich aus ausgewiesenen Fachvertretern der psychologischen Diagnostik zusammen. Die Kommission hat zum Ziel, Empfehlungen für die Studierendenauswahl im Fach Psychologie zu erarbeiten und Vorschläge zu entwickeln, wie die Psychologie ihre fachliche Expertise bei Selektions- und Plazierungsentscheidungen fachübergreifend einbringen kann.

Das mittel- und langfristige Ziel der Kommission ist es, Studierfähigkeitstests zu entwickeln, die sowohl allgemeine, fachunspezifische Komponenten, als auch spezifische Module enthalten. Mit diesen Testmodulen sollte erfasst werden, wie gut ein Bewerber oder eine Bewerberin den Anforderungen eines Studienfachs erfüllt. Die einzelnen Hochschulen hätten dann die Möglichkeit, sich aus dem multidimensionalen Test ihre spezifischen Profile zu erstellen (DGPs, 2005).

Für das Wintersemester 2005/06 wird von der Kommission für die anstehende Auswahl von Studierenden für das Fach Psychologie die Heranziehung der Kriterien Durchschnittsnote der Hochschulzugangsberechtigung und / oder gewichtete Einzelnoten empfohlen. Ausdrücklich abgeraten wird dagegen von Auswahlgesprächen (s. Passage zu möglichen Prädiktoren des Studienerfolgs in diesem Kapitel). Als weitere Kriterien werden standardisierte Verfahren empfohlen, die Aspekte der fachspezifischen Studierfähigkeit, wie mathematische Grundkenntnisse oder Textverständnis, prüfen (DGPs, 2005). Die potentiellen Verfahrensentwickler werden zudem angehalten, „Self-Assessment“-Tools für die Beratung von Studierenden bereitzustellen.

Generell empfiehlt die Kommission eine Kombination aus Schulnoten und dem Ergebnis eines Studierfähigkeitstests. Solch eine Kombination soll übereinstimmend mit der empirischen Befundlage den besten Prädiktor für den Studienerfolg darstellen (DGPs, 2004).

Pilotprojekt: Studierendenauswahl im Diplomstudiengang Psychologie an der J. W. Goethe-Universität

In Zusammenfassung der aufgeführten Überlegungen wurde zur Durchführung und Evaluation der Studierendenauswahl im Diplomstudiengang Psychologie an der J. W. Goethe-Universität folgendes Rahmenmodell entwickelt (s. Abbildung 2).

Anforderungsanalyse

Ausgangspunkt des Modells stellt eine Anforderungsanalyse dar (s. z. B. Westhoff et al., 2004), in der die zentralen Anforderungen des Studiengangs ermittelt werden sollten. Die Erstellung des Anforderungsprofils sollte sich einerseits auf die Critical Incident Technique (CIT) von Flangan (1954) stützen; andererseits sollten aber auch die Ergebnisse aus der Absolventenstudie sowie Überlegungen zur Profilbildung ("Zielvereinbarungen", Fachbereich Psychologie und Sportwissenschaften, 2005) und zum geforderten Leistungsniveau der Studierenden Berücksichtigung finden. Der zentrale Gedanke ist hierbei, eine möglichst gute Übereinstimmung zwischen den Erwartungen, Bedürfnissen und Wertorientierungen der Studierenden mit den Gegebenheiten und Chancen, welche der Studiengang bietet, also zwischen den Forderungen der Hochschule und den Potentialen der Bewerber (Person-Job-Fit; Amelang, 1997) zu erzielen.

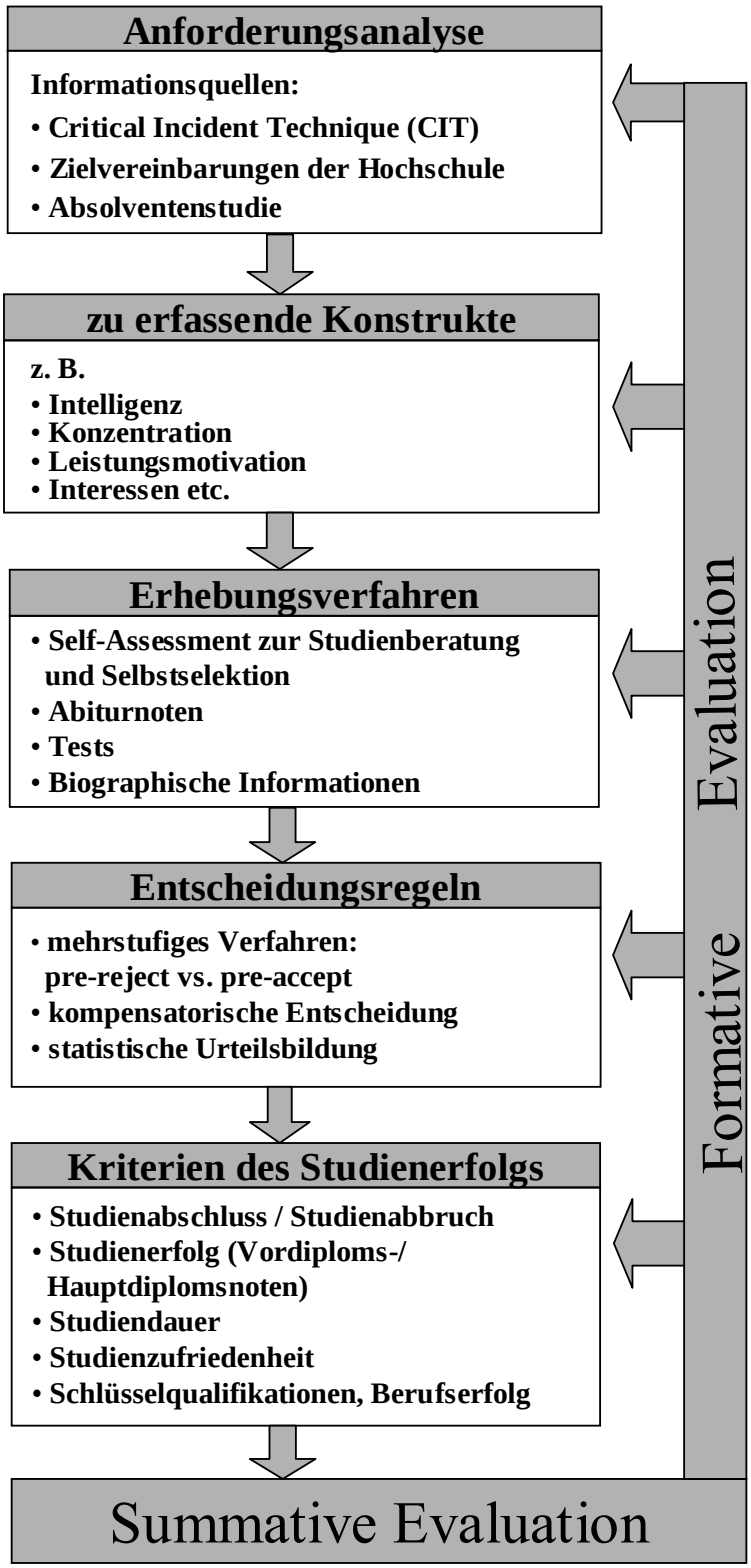


Abbildung 2: Rahmenmodell zur Durchführung und Evaluation der Studierendenauswahl

Zu erfassende Konstrukte

Die Anforderungsanalyse ermöglicht eine Benennung der zu erfassenden Konstrukte. Als studienrelevante Merkmale könnten sich bspw. Intelligenz, Konzentration, Leistungsmotivation und Interessen ergeben.

Erhebungsverfahren

Die konkreten Erhebungsverfahren können nur unter Berücksichtigung der rechtlichen Rahmenbedingungen (Bundes- und Landesgesetz, Universitätssatzung) festgelegt werden. Dabei sind die jeweiligen Bedingungen für die Realisierung zu definieren, z. B.: ob die Durchführung der Verfahren und der Studierendenauswahl in den Fachbereichen oder zentralisiert vorgenommen wird, ob Computer zum Einsatz kommen, wie die Finanzierung erfolgt etc. Vor der Selektion durch die Hochschule sollte den Bewerbern die Möglichkeit einer Beratung und ggf. einer Selbstselektion gegeben werden. Unter Berücksichtigung des gegenwärtig weit verbreiteten Zugangs zum Internet, bietet sich hier die Möglichkeit, ein webbasiertes Tool bereitzustellen. Für das Auswahlverfahren selbst kommen als valide Informationsquellen insbesondere Abiturnoten, Tests und biografische Informationen in Frage, deren Erhebung relativ zeit- und kostengünstig ist.

Entscheidungsregeln

Entscheidungsregeln können nach mehrstufigen, kompensatorischen und statistischen Entscheidungen differenziert werden. Bei den mehrstufigen Entscheidungsregeln sind vor allem zwei alternative Strategien von Interesse: zum einen die Vorweg-Ablehnungs-Strategie (pre-reject), bei der die Bewerber bei Nicht-Erfüllen eines ersten Auswahlkriteriums endgültig abgelehnt werden, und nur die Verbleibenden zu den weiteren Auswahlkriterien zugelassen werden. Ein Beispiel dafür wäre, dass zunächst nach den Abiturnoten "gesiebt" wird und nur Bewerber mit einem Abiturnotenschnitt von 2,5 in die engere Auswahl kommen; zum anderen die Vorweg-Annahme-Strategie (pre-accept), nach der die Ergebnisse des ersten Auswahlkriteriums dazu genutzt werden, eine Anzahl von Bewerbern aufzunehmen, die anderen müssen dann weitere Verfahren absolvieren. Die zweite Lösung ist in der aktuellen Hochschulzulassung insoweit gesetzlich vorgesehen, als 40% der Studienplätze bereits durch die ZVS vergeben werden (20% an die Abiturbesten und 20% nach der Wartezeit). Die Freiheit bei der Strategiefestlegung besteht lediglich in Bezug auf die "restlichen" 60% der Bewerber.

Wenn die Auswahlentscheidung aufgrund mehrer Kriterien erfolgen soll, stellt sich die Frage nach ihrer Kombination. Sinnvoll wäre die Verwendung sowohl der konjunktiven, als auch der kompensatorischen Strategie, z. B. indem Ergebnisse der Abiturnoten und des Studierfähigkeitstests kompensatorisch untereinander ausgeglichen werden dürfen, aber zugleich das notwendige Maß an Englischkenntnissen sichergestellt sein muss. Die Verrechnung der Ergebnisse der zur Auswahl eingesetzten Verfahren miteinander sollte

dann nach einem zuvor festgelegten Algorithmus (im Sinne einer statistischen Urteilsbildung) vorgenommen werden.

Kriterien des Studienerfolgs

Für eine summative Bewertung des Auswahlverfahrens wird eine Festlegung der Studienerfolgskriterien benötigt. Die in Frage kommenden Kriterien sind: Studienabschluss bzw. Studienabbruch, Studienabschlussnoten, Studiendauer, Studienzufriedenheit, allgemeine berufsqualifizierende Kompetenzen sowie Berufserfolg (s. Abschnitt zu Kriterien des Studienerfolgs in diesem Kapitel).

In Anlehnung an das vorgestellte Modell wird das Auswahlverfahren im Fachbereich Psychologie an der J. W. Goethe-Universität gestaltet und evaluativ begleitet.

Formative Evaluation der Studierendenauswahl im Diplomstudiengang Psychologie an der J. W. Goethe-Universität

Im Diplomstudiengang Psychologie der J. W. Goethe-Universität wird eine formative Evaluation der Studierendenauswahl durchgeführt. Diese besteht aus mehreren Schritten und hat zum Ziel, das bestmögliche, an die Anforderungen und Gegebenheiten des Studiengangs angepasste Verfahren zu erstellen.

Bereits nach Abschluss jeder Phase des Verfahrens sollte eine systematische Analyse der Zwischenergebnisse erfolgen, um sie im Sinne einer formativen Evaluation zu bewerten und ggf. zu optimieren.

Den Anfang bildete die Analyse des Zusammenhanges von Studienerfolg mit den Abiturdurchschnittsnoten sowie Einzelnoten. Im Rahmen einer empirischen Studie (*"Prognostizierbarkeit des Studienerfolgs aus schulischen Leistungsdaten";* Moosbrugger, Jonkisz & Fucks, in Vorbereitung) wurde im Wintersemester 2004/05 eine differenzierte Analyse der Noten und Fächer des Abiturzeugnisses durchgeführt. Insgesamt wurde die komplette Population von N=345 Absolventen im Diplom-Studiengang Psychologie an der J. W. Goethe-Universität Frankfurt am Main (Prüfungsordnung vom 7. Juli 1993; Abschlussjahrgänge WS 1995/96 bis incl. SS 2004) erfasst. Der Median des Alters bei Studienbeginn beträgt 23, der Median der Studiendauer beträgt 14 Semester. 77,7% der Absolventen waren weiblich. Die beste Vorhersage konnte für das Kriterium Studienerfolg, operationalisiert als "Note im Hauptdiplom", erzielt werden. Mit Hilfe von schrittweisen Regressionsanalysen, Diskriminanzanalysen sowie AnswerTrees wurde untersucht, welche Informationen aus dem Abiturzeugnis zur Prognose des Studienerfolgs am besten geeignet sind. Das Modell (schrittweise Regressionsanalyse), in dem die meiste Varianz erklärt wird, zeigt vor dem Ersatz der missing data eine multiple Korrelation von R = .54 (Varianzerklärung 29%), bzw. nach dem Ersatz der missing data R = .58 (Varianzerklärung 34%); es enthält folgende UV's: durchschnittliche Note im Abitur, Anzahl der

Fremdsprachen, Anzahl der belegten Halbjahre in der 1. Fremdsprache seit der 5. Klasse, durchschnittliche Note in Englisch und durchschnittliche Note in Mathematik (berechnet jeweils als Durchschnitt aus den erzielten Punkten der belegten Halbjahre in Stufe 12 und 13 sowie ggf. den Punkten in der Abiturprüfung).

Als ein weiterer Schritt bei der Gestaltung des Auswahlverfahrens wird eine Anforderungsanalyse durchgeführt. In Anlehnung an die Critical Incident Technique von Flangan (1954), sollen alle Professorinnen und Professoren, alle wissenschaftlichen Mitarbeiterinnen und Mitarbeiter sowie Vertreter der Studierenden nach typischen und erfolgskritischen Situationen im Psychologiestudium in Frankfurt befragt werden. Aus diesen Daten werden anschließend kognitive und nicht-kognitive Anforderungsmerkmale des Studienfachs abgeleitet und ein Anforderungsprofil des Studiengangs erstellt. Das Profil ermöglicht dann einen gezielten Einsatz psychologischer Testverfahren und eine Erstellung eines webbasierten SelfAssessment-Tools für Bewerber. Falls mehrere Prädiktoren (bspw. Durchschnittsnote der Hochschulzulassungsberechtigung und Intelligenztest und Motivationstest) erhoben werden, muss geklärt werden, welche Entscheidungsregeln Anwendung finden sollen.

Da eine summative Evaluation des Auswahlverfahrens erst ca. in 13 Semestern möglich sein würde (soviel beträgt die durchschnittliche Studiendauer in Frankfurt), wird eine erste Validierung des Verfahrens anhand von Ergebnissen in Pflichtklausuren und von Vordiplomsnoten der zum Wintersemester 2005/06 und zum Sommersemester 2006 zugelassenen Studierenden durchgeführt und auf das Verfahren rückgekoppelt.

Literatur

Amelang, M. (1997). Differentielle Aspekte der Hochschulzulassung: Probleme, Befunde, Lösungen. In T. Herrmann (Hrsg.), *Hochschulentwicklung - Aufgaben und Chancen.* Heidelberg: Asanger.

Arnhold, N. & Hachmeister, C.-D. (2004). Leitfaden für die Gestaltung von Auswahlverfahren an Hochschulen. Centrum für Hochschulentwicklung, Arbeitspapier Nr. 52. Verfügbar unter: http://www.che.de/downloads/Gestaltung_Auswahlverfahren_AP52.pdf [03.03.2005].

Baron-Boldt, J., Schuler, H. & Funke, U. (1988). Prädiktive Validität von Schulabschlussnoten: Eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie, 2*(2), 79-90.

Bundesministerium für Bildung und Forschung (2004). *Die Reform der Hochschulzulassung durch das 7. HRGÄndG.* Verfügbar unter: http://www.bmbf.de/de/2570.php [03.03.2005].

Bundestag (o. J.). *Grundgesetz*. Der Berliner Beauftragte für Datenschutz und Informationsfreiheit. Verfügbar unter: http://www.datenschutz-berlin.de/recht/de/gg/-gg1_de.htm#art12/ [03.03.2005].

Bundestag (2004). Siebtes Gesetz zur Änderung des Hochschulrahmengesetzes (7. HRGÄndG). *2298 Bundesgesetzblatt Jahrgang 2004 Teil I Nr. 47*. Bonn.

Deutsche Gesellschaft für Psychologie e.V. (2005). Stellungnahme der Deutschen Gesellschaft für Psychologie e.V. (DGPs). Zur Auswahl von Studierenden durch die Hochschulen (vom 22. November 2004). *Psychologische Rundschau, 56(2),* S. 153-154.

Deutsches Institut für Normierung (2002). *DIN 33430. Anforderungen an Verfahren und deren Einsatz bei berufsbezogener Eignungsdiagnostik.* Berlin: Beuth.

Educational Testing Service (2005). *Graduate Record Examinations (GRE)*. Verfügbar unter: http://www.gre.org/ [03.03.2005].

Educational Testing Service (2005). Scholastic Aptitude Test (SAT). Verfügbar unter: http://www.collegeboard.com/splash [03.03.2005].

Educational Testing Service (2005). *Test of English as a Foreign Language. TOEFL.* Verfügbar unter: http://www.ets.org/toefl/ [03.03.2005].

Eisenhuth, C. (2004). Determinanten der Studiendauer im Diplomstudiengang Psychologie der Johann Wolfgang Goethe-Universität. Ergebnisse einer Absolventenbefragung. Frankfurt am Main: Unveröffentlichte Diplomarbeit am Institut für Psychologie der J. W. Goethe-Universität.

Fachbereich Psychologie und Sportwissenschaften (2005). *Zielvereinbarungen zwischen den psychologischen Instituten des Fachbereichs Psychologie und Sportwissenschaften (05) und dem Präsidium der Johann Wolfgang Goethe-Universität Frankfurt am Main.*

Flangan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Gold, A. (1988). *Studienabbruch, Abbruchneigung und Studienerfolg: Vergleichende Bedingungsanalysen des Studienverlaufs.* Frankfurt am Main: Peter Lang.

Gold, A. & Souvignier, E. (1997). Examensleistung und Studienerleben bei Hochschulabsolventen. *Zeitschrift für Pädagogische Psychologie, 11*(1), 53-63.

Institut für Test- und Begabungsforschung (Hrsg.) (1986). Der neue TMS. Originalversion des Tests für medizinische Studiengänge im besonderen Auswahlverfahren. Göttingen: Hogrefe.

Moosbrugger, H. & Eisenhuth, C. (2004). FPAF. Frankfurter-Psychologie-Absolventen-Fragebogen. In H. Moosbrugger, D. Frank & W. Rauch (Hrsg.), *Qualitätssicherung*

im Bildungswesen. Riezlern-Reader XIII (Arbeiten aus dem Institut für Psychologie der J. W. Goethe-Universität, Heft 3/2004; S. 135-156). Frankfurt am Main: Institut für Psychologie der J. W. Goethe-Universität.

Moosbrugger, H., Jonkisz, E. & Fucks, S. (in Vorbereitung). *Prognostizierbarkeit des Studienerfolgs aus schulischen Leistungsdaten.* (Arbeiten aus dem Institut für Psychologie der J. W. Goethe-Universität). Frankfurt am Main: Institut für Psychologie der J. W. Goethe-Universität.

Moosbrugger, H., Reiß, S. & Eisenhuth, C. (2005). Determinanten von Studiendauer und Studienerfolg im Diplomstudiengang Psychologie. Eine Absolventenstudie. *Zeitschrift für. Evaluation* (in Vorbereitung).

Reiss, S & Moosbrugger, H. (2004). *Prädiktoren von Studiendauer und Studienerfolg: Ergebnisse einer Absolventenbefragung im Diplom-Studiengang Psychologie der Johann Wolfgang Goethe Universität Frankfurt (WS 1995/96 – SS 2002)* (Arbeiten aus dem Institut für Psychologie der J. W. Goethe-Universität, Heft 4/2004). Frankfurt am Main: Institut für Psychologie der J. W. Goethe-Universität.

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.

Schneller, K. & Schneider, W. (2005). Bundesweite Befragung der Absolventinnen und Absolventen des Jahres 2003 im Studiengang Psychologie. *Psychologische Rundschau, 56(2),* S. 159-175.

Tent, L. (1998). Zensuren. In D. H. Rost (Hrsg.), *Handbuch Pädagogische Psychologie* (S. 580-584). Weinheim: PVU.

Trost, G., Blum, F., Fay, E., Klieme, E., Maichle, U., Meyer, M. & Nauels, H.-U. (1998). Evaluation des Tests für medizinische Studiengänge (TMS): Synopse der Ergebnisse. Bonn: Institut für Test- und Begabungsforschung.

Westhoff, K., Hellfritsch, L. J., Hornke, L. F., Kubinger, K. D., Lang, F., Moosbrugger, H., Püschel, A. & Reimann, G. (Hrsg.) (2004). *Grundwissen für die berufsbezogene Eignungsdiagnostik nach DIN 33430.* Lengerich: Pabst Science Publishers.

Zuordnungs- und Klassifikationsstrategien in der Eignungsbeurteilung

Heike Mir

Einleitung

Wie viele andere Begriffe in der Psychologie so haben auch diejenigen von Diagnose und Diagnostik ihre Wurzeln im Griechischen, wo das Verb „ diagignostikein“ eine kognitive Funktion mit den Bedeutungen „gründlich kennenlernen“, „entscheiden“ und „beschließen“ bezeichnet (Amelang & Zielinski, 2002). Jäger und Petermann (1992) fassen psychologische Diagnostik als System von Regeln, Anleitungen und Algorithmen zur Bereitstellung von Instrumenten auf, mit deren Hilfe sowohl psychologisch relevante Charakteristika von Merkmalsträgern gewonnen als auch die erhobenen Daten zu einem diagnostischen Urteil integriert werden sollen, und zwar mit dem Ziel einer Vorbereitung von Entscheidungen sowie Prognosen und deren Evaluation.

Nach Amelang und Zielinski (2002) müssen auf der Basis der erhobenen diagnostischen Informationen Entscheidungen über anstehende Fragen gefällt werden, etwa der Art, ob ein Bewerber einen angestrebten Studienplatz erhält oder welche Schüler zweckmäßigerweise welchen Unterrichtseinheiten zugeordnet werden. In einem allgemeinen Sinn gehören der Studienplatz und Unterrichtseinheiten in die Kategorie von Interventionen, d.h. Maßnahmen, die aus den verschiedenen Gründen eingeleitet werden. Sie setzen an Diagnostische Feststellungen an, mit dem Ziel, Veränderungen auf organisatorischer oder individueller Ebene herbeizuführen. Nachfolgend sollen die Probleme, Fehler und Lösungsmöglichkeiten erörtert werden, die sich bei der Zuordnung von Diagnostischen Daten zu Interventionen ergeben. Diesbezüglich wird im Text insbesondere Bezug auf den Artikel von Rindermann und Oubaid (1999) genommen, indem es um die Auswahl von Studienanfängern durch Universitäten geht. Bei der Auswahl von Studienanfängern erhalten deutsche Hochschulen zukünftig einen größeren Spielraum. Allerdings sehen sich die Hochschulen mit der Aufgabe konfrontiert, ein hinreichend objektives, zuverlässiges, valides sowie faires und ökonomisches Bewerberauswahlverfahren zu entwickeln. Als adäquate Lösung wird ein mehrstufiges flexibles Auswahlmodell (ATIM) unter Einbezug von Abiturdurchschnittsnoten, Fachnoten, Eignungstests und Aufnahmegesprächen vorgeschlagen, das fachspezifische Gewichtungen der einzelnen Prädiktoren ermöglicht (Rindermann & Oubaid, 1999). Im folgenden Text sollen Zuordnungs- und

Klassifikationsstrategien aufgezeigt werden, die hilfreich dabei sein können, ein geeignetes Bewerberauswahlverfahren zu entwickeln.

Arten diagnostischer Entscheidungen

Klassifiziert werden diagnostische Entscheidungen nach einem Raster, das auf das nachgerade epochale Buch von Cronbach und Gleser (1965, zitiert nach Amelang & Zielinski,2002) zurückgeht (Tab.1). Aus der Kombination aller Klassifikationskriterien mit allen anderen resultieren $2^6 = 64$ verschiedene Arten von Diagnostischen Entscheidungen. Im folgenden Abschnitt erfolgt eine Beschränkung auf die häufiger vorkommenden Konstellationen.

Tabelle 1. Arten diagnostischer Entscheidungen (Amelang & Zielinski,2002)

1. Nutzen der Entscheidung geht zugunsten	Institution	vs.	Individuum
2. Annahme	festgelegt	vs.	variabel
3. Behandlungen	singulär	vs.	multipel
4. Möglichkeiten von Ablehnungen	ja	vs.	nein
5. Informationsdimensionen	univariat	vs.	multivariat
6. Entscheidungen	teminal	vs.	investigatorisch

Zu 1.

Eine Entscheidung ist von institutioneller Art, wenn eine Organisation (z.B. eine Fortbildungsanstalt) nach einem standardisierten Vorgehen alle Personen in der gleichen Weise einem Verfahren unterzieht. So müssen z.B. alle Personen ein und denselben Test bearbeiten oder an einem Vorstellungsgespräch teilnehmen, dessen Ergebnisse dann für die „Behandlung“ relevant sind (Amelang & Zielinski, 2002). Eine Entscheidung institutioneller Art liegt z.B. auch bei der Auswahl von Studienanfängern durch die Universitäten vor.

Geht der Nutzen der Entscheidung zugunsten des Individuums sind die Verhältnisse nach Amelang und Zielinski (2002) ganz anders gelagert, denn in diesem Fall geht ein Individuum auf eine Institution zu (z.B. um Rat über die anstehende Berufswahl einzuholen) und dort je nach Biographie und Vorkenntnissen ein spezifisches Untersuchungsprogramm mit dem Ziel zusammengestellt wird, die beste Handlungsalternative für die nachfragende Person herauszufinden. Hierbei interessiert allein der individuelle Nutzen.

Zu 2.

Um festgelegte Annahmequoten handelt es sich dann, wenn z.B. nur eine bestimme Zahl von Ausbildungsplätzen zur Verfügung steht, denen die Interessenten oder Bewerber zugeordnet werden müssen. Nach Rindermann und Oubaid (1999) werden bei einem Quotenmodell die Studienplätze nach unterschiedlichen Kriterien vergeben. So kann ein Teil der Plätze durch Berücksichtigung von Schulleistungen, ein anderer nach der Kombination von Abiturnote und Fähigkeitstests und ein dritter nach dem Resultat eines Auswahlgesprächs vergeben werden. Der besondere Vorteil dieser Variante besteht in ihrer relativ ökonomischen Struktur, da nur ein Teil der Bewerber an Tests oder Auswahlgesprächen teilnimmt.

Hingegen ist bei variablen Annahmequoten wechselseitige Unabhängigkeit der Entscheidungen über die einzelnen Probanden gegeben (Amelang & Zielinski, 2002)

Zu 3.

Unter Behandlung verstehen Cronbach und Gleser (1965, zitiert nach Amelang & Zielinski, 2002), die Unterscheidung zwischen einstufigen und mehrstufigen (sequentiellen) Testungen. Im ersten Fall erfolgt die Zuordnung auf der Basis einer punktuell-einmaligen Diagnose, im letzten als Resultat eines gestuften Vorgehens in mehreren Schritten.

In der diagnostischen Praxis kommt aus Zeit- und Kostengründen das einstufige Vorgehen recht häufig vor (Abb.1). Hier unterscheiden Amelang und Zielinski (2002) die:

- „nichtsequentielle Batterie“: Die gesamte Batterie wird an alle Probanden vorgegeben, und es werden diejenigen ausgewählt (III), die in dem optimal gewichteten Summenwert die höchsten Scores erzielen
- „single screen“: Auf einem Test allein (Annahmebereich = II) fußen alle weiteren Entscheidungen.

Ein einstufiges multiples Modell sieht nach Rindermann und Oubaid (1999) die Kombination aller Auswahlkriterien in einer Stufe vor. Danach würden beispielsweise alle Bewerber nach ihrer Abiturdurchschnittsnote, der Leistung in einem Fähigkeitstest und einem Interview beurteilt.

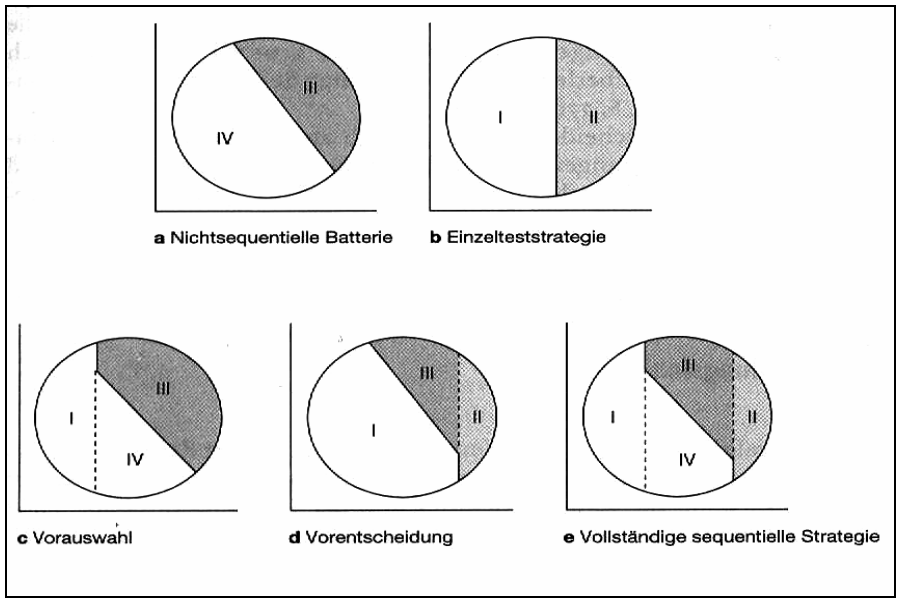


Abbildung 1 a-e. Zwei nichtsequentielle (a und b) und 3 sequentielle (c-e) Auswahlstrategien (Cronbach & Gleser, 1965, zitiert nach Amelang und Zielinski,2002).

Innerhalb des sequentiellen Vorgehens sind die folgenden 3 Grundmuster möglich (Abb.1):

- Vorauswahl-Strategie: Nach einem ersten Test werden alle Probanden, die einen bestimmten Score nicht erreichen, von weiteren Untersuchungen ausgeschlossen und zurückgewiesen (I). Die verbleibenden Probanden absolvieren weitere Verfahren. Die Entscheidung über Annahme (III) vs. Ablehnung (IV) wird aus der Kombination zwischen Erst- und Folgetests getroffen.
- Vorentscheidungs-Strategie: Nach einem ersten Teil werden alle Probanden, die einen bestimmten Trennwert überschreiten, bereits (terminal) akzeptiert (II). Mit den verbleibenden Probanden wird analog zur Vorauswahlstrategie verfahren.
- Vollständige sequentielle Strategie: Kombination der beiden vorgenannten Vorgehensweisen. Nach Maßgabe der Punktwerte in einem Test erfolgt eine Aufteilung aller Probanden in 3 Gruppen, eine, die (terminal) akzeptiert (II), eine andere, die definitiv abgewiesen (I) und eine dritte, die mit einem Folgetest untersucht wird.

Analog zum Thema, Auswahl von Studienanfängern durch Universitäten sieht das sequentielle Modell laut Rindermann und Oubaid (1999) die Auswahl in mehreren Stufen vor. So könnten beispielsweise nur die Abiturbesten zum Fähigkeitstest und hier wiederum nur die besten zum Interview zugelassen werden. In diesem Falle wären jedoch schwache Schulleistungen nicht mehr durch überzeugende Ergebnisse im Auswahlverfahren ausgleichbar. In einer anderen Variante könnte am Anfang die Zulassung eines Teils der Studienbewerber aufgrund der Abiturnote und die weitere Zulassung eines Teils nach Kombination von Abiturdurchschnittsnote und Fähigkeitstest stehen. In einem letzten

Schritt würden von den verbleibenden Personen mit unklarer Prognose mittels eines Auswahlgesprächs die restlichen Studienanfänger ausgewählt.

Dieses sequentielle Modell mit vorgegebenen Quoten verbindet die Vorzüge der beiden anderen Ansätze und bietet bei ökonomischen Vorteilen – das ökonomischste Verfahren Abiturnote steht am Anfang und nur ein Teil der Bewerber nimmt am Test und Auswahlgespräch teil – eine hohe Fairness und geringe Belastung für Institution und Bewerber . Konkretisiert ließe sich diese sequentielle Auswahlprozedur – das Abitur-Test-Interview-Modell (ATIM) – folgendermaßen in Tabelle 2 darstellen (Rindermann & Oubaid, 1999).

Tabelle 2: Schematischer Aufbau der sequentiellen Prozedur des ATIM (Rindermann & Oubaid,1999)

Kriterien	Verfahren	Studienplätze
Leistung, Fähigkeiten (indirekt Arbeitshaltung etc.)	Vorauswahl **Abiturdurchschnittsnote** (evtl. besondere Berücksichtigung studiennaher Fachnoten)	Vergabe von 40 % der Plätze an Abiturbeste
spezifische und allgemeine Fähigkeiten und Eignungen, eventuell Kenntnisse	2. Stufe **Kognitive Fähigkeitstests** (studienfachbezogene Tests, allgemeine kognitive Fähigkeiten)	Vergabe von 40 % der Plätze an Abitur- und Testbeste (Kombination), Berücksichtigung von Fachnoten
Studienmotivation, Interessen, soziale Kompetenz, Engagement	3. Stufe **Auswahlgespräch** (strukturiertes Interview, Leitfaden)	Vergabe von 20 % der Plätze, Verhältnis 2:1 oder 3:1 Interviewte/Plätze

Je nach Anforderungscharakteristika spezifischer Studiengänge und der Bedeutung, die man der Auswahl geeigneter Bewerber zumisst, ließen sich die Quoten innerhalb des Abitur-Test-Interview-Modells verschieben. Die relative Überlegenheit von sequentiellen zu nichtsequentiellen Strategien ist nach Amelang und Zielinski (2002) bei institutionellen Entscheidungen an Nutzenüberlegungen gekoppelt, d.h. die Gewinne, die eine Organisation daraus erwirtschaftet, dass auf der Basis von diagnostischen Untersuchungen die Bestgeeigneten identifiziert werden, im Vergleich zu den Kosten, die eben diese Testungen verursachen.

Zu 4.

Sind Ablehnungen aufgrund von Testungen möglich, liegt die klassische Struktur von Selektionsparadigmen vor. Verbleiben hingegen alle Probanden im System und werden infolge der Diagnosestellung nur horizontal oder vertikal zu spezifischen Interventionen „verschoben“, spricht man von Platzierung (Klassifikation)(Abb.2).

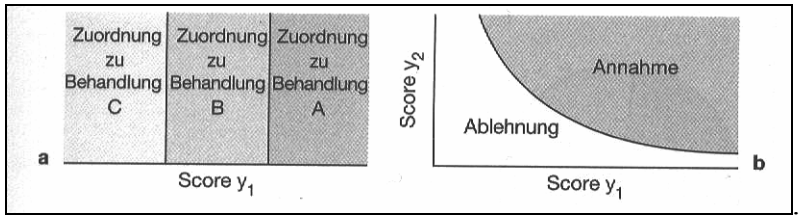


Abbildung 2. a. Platzierung; b. Selektion (aus Cronbach & Gleser,1965; zitiert nach Amelang und Zielinski, 2002)

Zu 5.

Die diagnostische Information kann sich auf eine Dimension beschränken (z.B. die Abiturnote), also univariat vorliegen, oder aus mehreren Dimensionen stammen und somit multivariat beschaffen sein. Meist werden zur Erhöhung der Validität und damit auch der Entscheidungssicherheit mehrere Prädiktoren herangezogen, weil damit verschiedene Facetten des Kriteriums abgedeckt werden können (Amelang & Zielinski, 2002).

Zu 6.

Wird auf der Basis der diagnostischen Information ein Proband einer Behandlung zugeführt, in der er mehr oder weniger lange verbleibt (z.B. Aufnahme in ein Ausbildungsprogramm), handelt es sich um eine terminale Entscheidung. Mit der Zuweisung ist die diagnostische Aufgabe abgeschlossen. Soll die Maßnahme, der eine Person als Ergebnis diagnostischer Datensammlung zugeordnet wird, hingegen nur vorläufig sein (z.B. eine Anstellung auf Probe), sprechen wir von einer investigatorischen Entscheidung.

Kompensatorische und konjunktive Entscheidungsstrategien

Die lineare Kombination von Prädiktionswerten zu einem Rechenmaß, das eine maximale Korrelation mit dem jeweiligen Kriterium gewährleistet, impliziert ein sog. kompensatorisches Modell. Das heißt, ein und derselbe Prädiktionswert kann durch ganz verschiedene Merkmalskonfigurationen in den Einzeltests erreicht werden (Amelang & Zielinski, 2002). Kompensatorische Modelle liegen laut Amelang und Zielinski (2002) der diagnostischen Praxis sehr häufig zugrunde. Etwa kann das Ziel der Versetzung in die nächste Schulklasse auch bei starken Defiziten in bestimmten Fächern erreicht werden, wenn diese durch besonders gute Leistungen in anderen ausgeglichen werden.

Eine andere Entscheidungsstrategie beschreibt das „Oder-Konzept“. Dort ist es nach Amelang und Zielinski (2002) nicht notwendig, die Summe aus Teilkompetenzen zu bilden, sondern es genügen entsprechend hohe Punktwerte in einem der Prädiktoren. Eine solche

Auswahlstrategie liegt dann nahe, wenn die durch das Kriterium geforderte Leistung entweder auf die eine oder andere Weise erbracht werden kann.

Kompensatorische Strategien sind immer dort dysfunktional, wo in jedem Teilbereich bestimmte Mindestleistungen unabdingbar vorliegen müssen, um eine Tätigkeit erfolgreich ausführen zu können. Beispielsweise kann ein Chirurg nicht mangelnde feinmotorische Kompetenz durch Intelligenz kompensieren. Hier besteht die Forderung nach Leistungen in dem einen und dem anderen Bereich, weshalb diese Modelle auch konjunktive bzw. „Und-Strategien" heißen (Amelang & Zielinski, 2002). In Bezug auf die Studierendenauswahl durch die Universitäten stellt sich die Frage welches der beiden Strategien gewählt werden sollte.

Ein kompensatorisches und ein konjunktives Modell sind nach Amelang und Zielinski (2002) in Abbildung 3a und b für den Fall graphisch veranschaulicht, dass der für die Zulassung kritische Testtrennwert mit $z_{\hat{y}} = -1$ festgelegt worden wäre.

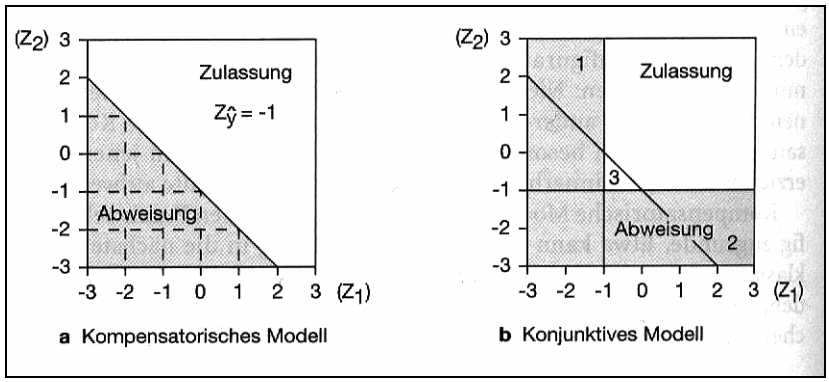


Abbildung 3a und b. Entscheidungsstrategien: a kompensatorische, b konjunktive. (Aus Wieczerkowski & Oeveste, 1982; zitiert nach Amelang & Zielinski, 2002).

Nach der kompensatorischen Strategie (Abb.3a) fallen alle Probanden in die Kategorie „Zulassung", bei denen die Kombination aus Z_1 und Z_2 mindestens den Wert $z_{\hat{y}} = -1$ ergibt (also $Z_1 = +2$, $Z_2 = -3$; $Z_1 = +1$, $Z_2 = -2$ usw.). Da ein z – Wert von −1 einem Prozentrang von 16 entspricht, gehören 84% aller Probanden in die Kategorie „Zulassung", deren Grenze durch die schräge Gerade in Abbildung 3a markiert wird.

Dem konjunktiven Modell (Abb.3b) zufolge ist der kritische Trennwert in jeder der beiden Variablen bei z = −1 angesetzt worden. Daraus resultiert ein insgesamt konservativeres Vorgehen, d.h. die Anforderungen sind höher, um in die Kategorie der Zugelassenen zu gelangen. Dementsprechend fallen nun die mit 1 und 2 bezeichneten Segmente – im Gegensatz zur kompensatorischen Strategie – unter die Abgelehnten. Gleichwohl gäbe es auch einige Probanden, die unter den gegebenen Randbedingungen unter der konjunktiven, aber nicht unter der kompensatorischen Strategie zugelassen werden (Amelang &Zielinski,

2002). In Bezug auf die Studierendenauswahl durch Universitäten, stellt sich die Frage, welcher Strategie der Vorzug gewährt werden sollte. Auch dieser Punkt soll Anlass zur Diskussion bieten.

Klassifikation und Selektion

Klassifikation

Klassifikation bezeichnet einen diagnostischen Prozess, in dem „einer Person bestimmte sozial definierte Dispositionen zugeschrieben werden. Ziel der Persönlichkeitsdiagnostik ist es dann, die Berechtigung dieser Zuschreibungen zu überprüfen.“ Beispielsweise kann ein Zehnjähriger bei der Wahl der „richtigen“ weiterführenden Schule Hilfe suchen (Fisseni, 2004).

Allgemein lässt sich die diagnostische Aufgabe bei einer Klassifikation als Frage formulieren: In welche Merkmalsklasse gehört der Proband (Tab.3)? Um die Aufgabe der Klassifikation zu lösen muss der Diagnostiker laut Fisseni (2004) drei Festlegungen treffen:

- Er muss die Klassen genau definieren.
- Er muss Kriterien angeben, die bestimmte Leistungen abgrenzen.
- Er muss Entscheidungsregeln formulieren, die besagen, bei welcher Leistung ein Proband einer Klasse zugewiesen wird.

Tabelle 3: Klassifikationsaufgabe – schematisiert. (Aus Fisseni, 2004)

Gegeben	In welche	> In Klasse A?
Ist	Merkmalsklasse	> In Klasse B?
Proband 1	gehört Proband 1	> In Klasse C?

Selektion

Bei dem anderen Ansatz diagnostischer Aufgaben, der Selektion, ist eine „Merkmalsklasse“ vorgegeben, für die der „richtige“ Proband zu suchen ist. Beispiel: Ein Unternehmen A bietet Lehrstellen an, um die sich „begabte“ Sechzehnjährige bewerben können. Laut Fisseni (2004) lässt sich die diagnostische Aufgabe in eine Frage fassen, die komplementär zu der des ersten Ansatzes lautet: Welcher Proband besitzt die Merkmale, die ihn für die angebotene „Stelle“ als geeignet ausweisen? Selektion schließt die Möglichkeit der Ablehnung ein. Beispiele vordefinierter „ Merkmalsklassen“, für die eine Person zu suchen ist, sind Stellen, die ein Betrieb anbietet, oder Studienplätze, die zur Verfügung stehen (Tab.4).

Tabelle 4: Selektionsaufgabe – schematisiert. (Aus Fisseni, 2004)

Gegeben	Welcher Proband	> Proband 1?
Ist	ist für Stelle A am	> Proband 2?
Stelle A	besten geeignet?	> Proband 3?

Drei Faktoren sollte der Diagnostiker bei der Selektion berücksichtigen: Die Basisrate, die Selektionsrate und die Validität der eingesetzten Verfahren (Fisseni, 2004).

Die Basisrate (BR) bezeichnet die Häufigkeit, mit der ein bestimmtes Merkmal in einer Bestimmten Gruppe auftritt. Geht man von einem konkreten Beispiel aus – etwa einer Stelle, um die sich Bewerber bemühen – dann definiert sich die Basisrate als Verhältnis der Zahl geeigneter Bewerber (N_c) zur Bewerbergesamtzahl (N_g); in einer Formel besagt dies:

$BR = N_c/N_g$

Dabei geht es um die tatsächlich Geeigneten, eine Zahl, die in der Regel nicht bekannt und kaum exakt bestimmbar ist. Das Problem besteht darin, die Zahl dieser geeigneten Personen zu schätzen (Fisseni, 2004).

Die Selektionsrate (SR) bezeichnet nach Fisseni (2004) den relativen Anteil der auszulesenden Personen an der Gesamtzahl der Personen, die sich der Auslese stellen. Geht man wieder von dem Beispiel eines Stellenangebotes aus, dann definiert sich die Selektionsrate als Verhältnis der Zahl offener Stellen (N_o) zur Bewerbergesamtzahl (N_g), in einer Formel umschrieben besagt dies:

$SR = N_o/N_g$

Beispiel:

- 123 Kandidaten bewerben sich um 13 Sekretärsstellen. Es sei angenommen, dass unter den Probanden 25 Personen für die Stelle geeignet sind. Die Basisrate bestimmt sich dann wie folgt; $BR = N_c/N_g = 25/123 = 0.20$. Die Basisrate beträgt 0.20; das heißt, 20 Prozent der 123 Kandidaten sind „geeignete“ Bewerber.
- Die Selektionsrate bestimmt sich wie folgt: $SR = N_o/N_g = 13/123 \approx 0.11$. Die Selektionsrate beträgt 0.11; das heißt, etwa 11 Prozent der Bewerber können eine Stelle bekommen.
- In dem Beispiel ist die Basisrate (0.20) größer als die Selektionsrate (0.11). Die Aufgabe besteht nun darin, aus der Zahl der 123 Kandidaten dreizehn der 25 „Geeigneten“ zu identifizieren. Bei diesem Anliegen kann die Beachtung der Validität eines Verfahrens weiterführen.

Taylor und Russell (1939, zitiert nach Fisseni, 2004) haben ein Tafelwerk vorgelegt, das es ermöglicht, den Zusammenhang zwischen Basis- und Selektionsrate sowie der Validität eines Verfahrens in Form einer Trefferquote zu schätzen.

- Bezogen auf unser Beispiel würde dies bei einer Basisrate von 0.20 bedeuten: Unter der Gesamtzahl der 123 Bewerber sind 20% tatsächlich geeignet. Demnach kann man bei einer Selektion nach Zufall, etwa nach Losverfahren, mit 20% Treffern rechnen.
- Setzt man bei einer Basisrate von 0.20 und einer Selektionsrate von 0.11 ein Instrument ein, das eine Validität von 0.65 besitzt, dann ist nach den Taylor-Russel-Tafeln eine Trefferquote von 0.64 zu erwarten.
- Das bedeutet: Im Vergleich zu einer Zufallsauswahl mit 20% erwarteter Treffer ergibt sich ein Zugewinn von 44 Prozentpunkten.

Entscheidungsfehler

Die zentrale Aufgabe von Zuordnungsstrategien besteht darin, Fehler bei der Klassenzuordnung zu vermeiden. Derartige Fehler liegen immer dann vor, wenn die Zuordnung aufgrund der Prädiktorvariablen nicht mit der tatsächlichen Klassenzugehörigkeit übereinstimmt (Amelang & Zielinski, 2002).

Nach Fisseni (2004) ist eine Zuordnung richtig, wenn ein „Merkmalsträger“ jener Klasse zugewiesen wird, die das Merkmal repräsentiert, das er besitzt. Typisch sind zwei Fälle richtiger Klassifikation:

- Der Geeignete wird der Klasse der Geeigneten zugewiesen, man spricht von „wahren Positiven“. Beispiel: Begabte werden als begabt eingestuft.
- Der Ungeeignete wird der Klasse der Ungeeigneten zugewiesen, man spricht von „wahren Negativen“. Beispiel: Unbegabte werden eingestuft als unbegabt.

Falsch ist eine Zuordnung, wenn ein „Merkmalsträger“ einer Klasse zugeordnet wird, die ein Merkmal repräsentiert, das er nicht besitzt. Typisch sind zwei Arten von Fehlzuweisungen:

- Probanden, die geeignet sind, werden der Klasse der Ungeeigneten zugewiesen, man spricht von „falschen Negativen“. Es handelt sich um den so genannten β-Fehler. Beispiel: Begabte werden als unbegabt eingestuft.
- Probanden, die ungeeignet sind, werden der Klasse der Geeigneten zugeordnet, man spricht von „falschen Positive“. Es handelt sich um den so genannten α-Fehler. Beispiel: Unbegabte werden als begabt eingestuft.

In Bezug auf die Studierendenauswahl wäre es demnach aus Sicht der Universitäten sinnvoller, die „falsch Positiven“ zu minimieren, aus Sicht der Bewerber würde es jedoch umgekehrt aussehen.

Um die Fehler bei Zuordnungsverfahren gering zu halten, bieten sich mehrere Arten von Entscheidungsregeln an (nach Kallus & Janke, 1992; zitiert nach Amelang & Zielinski, 2002):

- das Neyman-Pearson-Kriterium erlaubt es, unterschiedliche Risiken von Fehlentscheidungen in die Klassenzuordnungsunterscheidung mit einzubeziehen. Sein Prinzip besteht darin, das Modell der statistischen Hypothesenprüfung auf die Klassenzuordnungsentscheidung anzuwenden. In Analogie zur Festlegung des kritischen Wertes der Teststatistik bei der Hypothesenprüfung wird das Entscheidungskriterium so verschoben, dass das Risiko für den Fehler erster Art unterhalb eines frei bestimmbaren Wertes liegt (z.B. $\alpha < 0{,}05$ oder 0,01 usw.).
- nach dem Minimax-Kriterium wird der maximale Zuordnungsfehler (betrachtet in allen Klassen) möglichst klein gehalten. Der Betrag des größten Zuordnungsfehlers aller Klassen/Kategorien/Gruppen ist am geringsten.
- Das Minimum-Loss-Kriterium minimiert die Zuordnungsfehler über alle Klassen hinweg. Dies kann im Vergleich zum Minimax-Kriterium bedeuten, dass eine Konstellation gewählt wird, bei der ein Zuordnungsfehler einer Kategorie/Klasse/Gruppe im Vergleich zu allen anderen relativ hoch ist.

Die Analyse von Zuordnungsfehlern setzt voraus, dass zuvor eine Zuordnung bereits stattgefunden hat. Diese kann sich verschiedener Methoden bedienen (Amelang & Zielinski, 2002):

- Zugehörigkeitswahrscheinlichkeiten: Auf der Basis von Wahrscheinlichkeitstafeln, wie sie Taylor und Russel (1939, zitiert nach Amelang & Zielinski, 2002) erarbeitet haben, erfolgt die Zuordnung zu derjenigen Klasse, der das Individuum nach Maßgabe der Ausprägung im Prädiktor mit der größten Wahrscheinlichkeit angehört. Dafür wird der Likelihood-Quotient herangezogen.
- Regressionstechniken: Durch Einsetzen der individuellen Prädiktionswerte in die für das anstehende Problem ermittelte Regressionsgleichung werden individuelle Kriteriumswerte ermittelt. Die Zuordnung zu den Kategorien erfolgt durch Differenzbildung mit kritischen Kriteriumswerten.
- Diskriminanzanalyse: Durch Einsetzen der individuellen Testwerte in die Diskriminanzfunktion resultiert ein Wert, der entweder größer, gleich oder kleiner ist als der kritische Diskriminationswert, der die Klassen voneinander trennt. Entsprechend kann anhand des individuellen Diskriminationswertes unmittelbar die Zuordnung zu einer der Gruppen vorgenommen werden.

- Ähnlichkeits- bzw. Distanzmaße: Häufig wird ein individuelles Testwerteprofil mit dem durchschnittlichen Profil verschiedener Gruppen von Personen (z.B. Schüler des sprachlichen oder mathematischen Zweiges) verglichen. Dafür stehen verschiedene Maße zur Verfügung, z.B. das Ähnlichkeitsmaß (Euklidische Distanz) von Osgood und Suci (1952, zitiert nach Amelang & Zielinski, 2002).

Unter anderem ist die von Lienert (1989, zitiert nach Amelang & Zielinski, 2002) adaptierte Cattellsche Formel gebräuchlich.

Nutzenerwägung

Nach Amelang und Zielinski (2002) werden institutionelle und individuelle Entscheidungen getroffen, da sich die jeweiligen Organisationen bzw. Personen im Fall richtiger Entscheidungen etwas davon versprechen, nicht zuletzt positive ökonomische Auswirkungen, also Gewinne, während bei falschen Entscheidungen die Gefahr von Verlusten droht.

Zur Einschätzung der Praktikabilität von Verfahren treten laut Schuler (2001) neben die Validität auch Aspekte wie die Kosten der Verfahrenskonstruktion oder -auswahl, der erforderliche organisatorische Aufwand bei der Durchführung und der zu erwartende Nutzen durch die Anwendung.

Cronbach und Gleser (1965, zitiert nach Fisseni, 2004) haben einen Algorithmus vorgeschlagen, der es erlauben soll, den Gesamtnutzen diagnostischer Untersuchungen zu schätzen, indem Gesamtnutzen und Gesamtkosten in einzelne Komponenten zerlegt werden. Die Formel lautet vereinfacht:

$$Nu_{(ges)} = Nu_{(e)} - Ko$$

Es bedeuten:

$Nu_{(ges)}$: Gesamtnutzen

$Nu_{(e)}$: Nutzen von Einzelkomponenten

Ko: Kosten

Demnach ergibt sich der Gesamtnutzen als Differenz zweier Größen: des Nutzens von Einzelkomponenten und des Kostenaufwandes. Der Nutzen von Einzelkomponenten ergibt sich als ein Produkt, in das eingehen

- der Informationsnutzen eines Scores (aus Test, Exploration oder projektivem Verfahren usw.)
- der Behandlungsnutzen, den ein Treatment liefert;
- der Ergebnisnutzen, der an einem Kriterium zu messen ist.

Die Kosten ergeben sich als Summe aller Elemente, die zu den Kosten beitragen (Fisseni, 2004).

Den Nutzen einer diagnostischen Untersuchung zu schätzen, wie Cronbach und Gleser vorschlagen, setzt nach Fisseni (2004) voraus, dass sich jede Größe genau bestimmen lässt und alle beteiligten Komponenten in gleichen Einheiten angegeben werden.

Doch gerade darin liegt die Schwierigkeit des Modells: wie soll beispielsweise der Nutzen geschätzt werden, der einer einzelnen Information zukommt? Wie der Nutzen, der einer Behandlung entspringt?

Neben dem Modell von Cronbach und Gleser (1965, zitiert nach Fisseni, 2004) werden in einzelnen Modellen Nutzen und Kosten unterschiedlich geschätzt, beispielsweise:

- durch Vergleich der Vorzüge verschiedener Methoden (Multi-Attributive-Utility-Theory: Maut)
- durch Angaben in Geld
- durch Befragung von Experten
- durch Befragung der Betroffenen

Die diagnostische Arbeit sollte sich nicht allein an den psychometrischen Gütekriterien diagnostisch-interventiver Verfahren orientieren, sondern auch an einer Schätzung des Nutzens, den eine Untersuchung für Proband und Mitwelt erbringt, und der Kosten, welche die Untersuchung verursacht (Fisseni, 2004).

Schlussbemerkung

Die vorgestellte Arbeit sollte zur Diskussion anregen, und hilfreich dabei sein wie eine Auswahlstrategie durch Universitäten gestaltet werden könnte. D.h. :

- Sollte nur die Abiturnote berücksichtigt werden (univariat) oder lieber mehrere Prädiktoren Berücksichtigung finden (multivariat)?
- Sollten einstufige oder sequentielle Modelle gewählt werden?
- Sollten kompensatorische oder konjunktive Strategien Verwendung finden?

- Welche Zuordnungsstrategie sollte gewählt werden, z.B. Zugehörigkeitswahrscheinlichkeiten, Regressionstechniken, Diskriminanzanalyse, Ähnlichkeits- bzw. Distanzmaße?
- Und nicht zuletzt die Frage, ob der Nutzen eines solchen Verfahrens im Verhältnis zu den Kosten steht, die eine solche Untersuchung mit sich bringt.

Literatur

Amelang, M. & Zielinski, W. (2002). Psychologische Diagnostik und Intervention *(3., korrigierte, aktualisierte und überarbeitete Auflage unter Mitarbeit von Thomas Fydrich und Helfried Moosbrugger.). Berlin: Springer.*

Fisseni, HJ. (2004). Lehrbuch der psychologischen Diagnostik. *Göttingen: Hogrefe.*

Jäger, R. S. & Petermann, F. (1992) Psychologische Diagnostik *(2. veränderte Aufl.). Weinheim: Psychologie Verlags Union.*

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.*

Schuler, H. (2001). Lehrbuch der Personalpsychologie*. Göttingen: Hogrefe.*

Das Allgemeine Lineare Modell – Lineare Regression

Dirk Frank

Einführung

Das Allgemeine Lineare Modell gehört zu den wichtigsten statistischen Verfahren der empirischen Wissenschaften. Zahlreiche Fragestellungen erfordern die mathematische Präzisierung und Modellierung der Zusammenhänge zwischen beobachteten Variablen, meist mit dem Zweck, einen erklärenden Prädiktionszusammenhang zwischen bedingenden, sog. „Prädiktorvariablen“ (unabhängigen Variablen) und bedingten sog. „Kriteriumsvariablen“ (abhängigen Variablen) zu modellieren. „Erklärung“ bedeutet in diesem Kontext nicht notwendigerweise die Herstellung eines kausalen Zusammenhanges zwischen Variablen (zum Konzept der Kausalität und den Bedingungen empirischer Kausalitätsbeweise vgl. Steyer, 1992). Vielfach dienen lineare Modelle der Darstellung empirischer Gesetzmäßigkeiten zwischen gemessenen bzw. beobachteten Größen und damit vor allem der Vorhersage von Zuständen/Ausprägungen bedingter Variablen bei vorliegender Kenntnis der Ausprägungen bedingender Variablen. Während vielen Fragestellungen der Allgemeinen Psychologie durch experimentelle Bedingungsvariation nachgegangen werden kann und die Erklärung der Verhaltensvariation zwischen den „Treatments“ von Interesse ist (die Variation zwischen den Probanden wird in der Regel als „Fehlervarianz“ vernachlässigt), fokussiert die Differentielle Psychologie in der Regel auf eben diese interindividuellen Unterschiede und versucht diese Variation auf Organismusvariablen. (bspw. Persönlichkeitsmerkmale) zurückzuführen. Traditionell erfolgt die Auswertung (labor)experimenteller Daten mit den Methoden der *Varianz- und Kovarianzanalyse* (Unterschiedshypothesen), dagegen werden die korrelativen, meist via Feldstudien („Quasi-Experimente“ oder Beobachtungen) operationalisierten Fragestellungen der Differentiellen Psychologie in der Regel mittels *Korrelations- und Regressionsmethoden* analysiert (vgl. Moosbrugger, 1994). Typische Fragestellungen sind bspw.:

- Der Studienerfolg soll prognostiziert werden aus prädiktiven Variablen wie Intelligenz, Alter und Geschlecht der Studierenden, sowie aus dem Einkommen und dem sozialen Status der Eltern.
- Modellierung der Abhängigkeit des Erfolgs psychotherapeutischer Maßnahmen von Alter, Geschlecht, Störungsbild und Persönlichkeitsvariablen des Patienten.
- Die Prognose künftigen Führungserfolgs in Abhängigkeit vom Antwortverhalten in berufsbezogenen psychometrischen Testbatterien.

Während in der Literatur lange Zeit eher die separate Beschreibung experimenteller und korrelativer Einzelverfahren vorherrschte, liefert das Allgemeine Lineare Modell einen übergreifenden Modellrahmen, der die kohärente und konsistente Darstellung dieser und anderer (bspw. Faktorenanalysen, logistische Regressionen) Analyseverfahren erlaubt (vgl. bspw. Cohen, 1968).

Klassifikation multivariater Verfahren

Häufig finden sich in der statistischen Literatur Einteilungen der verschiedenen Analysemethoden nach der Art der „Aufteilung" der Datenmatrix (vg. Tab.1), nach dem einem Analyseverfahren zugrunde liegenden „Erkenntnisinteresse" (vgl. Tab. 2) oder nach den zugrunde liegenden Skalenniveaus der Analysevariablen (vgl. Tab. 3). Darstellungen dieser Art fördern ebenfalls eher die Wahrnehmung der zahlreichen multivariaten Verfahren als „Einzelanwendungen", obwohl viele davon – jedoch auch wiederum nicht alle[2] - unter dem ALM-Dach formalisiert werden können.

Tab. 1: Klassifikation multivariater Analysemethoden nach der Partitionierung der Datenmatrix

Einteilungskriterium	Art des Analyseverfahrens	Verfahrensbeispiele
geteilte Variablenmenge	**Dependenzanalyse**	Regressionsanalyse Varianzanalyse Kovarianzanalyse Diskriminanzanalyse Kontingenzanalyse Logistische Regression
ungeteilte Variablenmenge	**Interdependenzanalyse**	Korrelationsanalyse Faktorenanalyse Clusteranalyse Multidimensionale Skalierung (MDS) Conjoint-Analyse Konfigurationsfrequenzanalyse (KFA)

[2] So lassen sich eine Reihe clusteranalytischer Verfahren nicht im Rahmen des ALM darstellen, eine Ausnahme innerhalb der Clusteranalysen bilden wiederum einige iterativ-partitionierende Verfahren, welche Varianzkriterien zur optimalen Klassifizierung verwenden (vgl. Moosbrugger & Frank, 1992).

Tab. 2:Klassifikation uni- und multivariater Analysemethoden nach der zugrunde liegenden Forschungsabsicht

Einteilungskriterium	Art des Analyseverfahrens	Verfahrensbeispiele
Strukturbeschreibend	**Deskriptive Datenanalyse**	Mittelwert Streuung Häufigkeiten
Primär strukturprüfend	**Konfirmatorische Datenanalyse**	Regressionsanalyse Varianzanalyse Diskriminanzanalyse logistische Regression Conjoint-Analyse Kontingenzanalyse „Kausalanalyse“ (Lineare Strukturgleichungsmodelle)
Primär strukturentdeckend	**Explorative Datenanalyse**	Faktorenanalyse Clusteranalyse Multidimensionale Skalierung (MDS)

Tab. 3: Klassifikation multivariater Analysemethoden nach dem Skalenniveau der Variablen

		Unabhängige Variable	
		Metrisches Skalenniveau	Nominales Skalenniveau
Abhängige Variable	Metrisches Skalenniveau	Regressionsanalyse	Varianzanalyse
	Nominales Skalenniveau	Diskriminanzanalyse Logistische Regression	Kontingenzanalyse

Die Modellgleichung des ALM

Die fundamentale Gleichung des Allgemeinen Linearen Modells drückt den mathematischen Zusammenhang zwischen Prädiktor- und Kriteriumsvariablen aus, indem die Ausprägungen in der Kriteriumsvariablen y als gewichtete Summe (Linearkombination) von Ausprägungen in einer oder mehreren Prädiktorvariablen dargestellt werden. Wir gehen im folgenden davon aus, dass von $i = 1,\ldots,n$ Probanden Messwerte in $j = 1,\ldots,m$ Prädiktorvariablen x_j mit den Gewichtszahlen β_j $(X_1,X_2,...,X_m)$ und einer Kriteriumsvariablen y_i vorliegen. Allgemein ist $n > m$. Als weiterer Prädiktor fungiert eine Prädiktionskonstante x_0, die über eine eigene Gewichtszahl β_0 verfügt.

Damit lautet die Gleichung auf Personenebene

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + e_i \quad (1)$$

für jeden Probanden i. In Matrizenschreibweise

$$\begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & ... & x_{1m} \\ . & . & ... & . \\ . & . & ... & . \\ . & . & ... & . \\ 1 & x_{n1} & ... & x_{nm} \end{pmatrix} * \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_m \end{pmatrix} + \begin{pmatrix} e_1 \\ . \\ . \\ . \\ e_n \end{pmatrix} \quad (2)$$

bzw. in allgemeiner Formulierung

$$y_i = \sum_{j=0}^{m} \beta_j x_{ij} + \varepsilon_i$$

Hierbei gibt der Fehlerterm ε_i die Größe der Differenz zwischen dem tatsächlichen und dem vorhergesagten Wert der Kriteriumsvariablen an.

Zentrale Eigenschaften des Allgemeinen Linearen Modells sind dabei:

- Additivität: Die einzelnen gewichteten Einflussgrößen addieren sich auf zum Wert der AV,
- Linearität: Das Modell ist linear in den Parametern, erlaubt jedoch die Modellierung kurvi-linearer Zusammenhänge ($\beta_2 x^2$, moderierte Regression $\beta_3 x_1 x_2$),
- Kompensation: Niedrige Werte in Prädiktoren können durch hohe Werte in anderen Prädiktoren kompensiert werden,
- Personeninvarianz: Die Gewichtung der Prädiktoren ist personunabhängig konstant.

Modell der Einfachregression

Bei Vorliegen einer einfachen Prädiktor-Kriteriumsbeziehung lautet das Schätzmodell

$$\hat{y}_i = \beta_0 + \beta_1 \cdot x_i$$

Bei einer mittelwertszentrieren Messung erfolgt die Schätzung über Abweichungswerte, die Regressionsgerade geht dabei durch den Nullpunkt des Koordinatensystems und die Regressionskonstante entfällt:

$$\hat{y}_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

Bei weiterer Standardisierung (Berechnung über z-Werte) entspricht β_1 dem Korrelationskoeffizienten.

Modell der multiplen Regression

$$\hat{y}_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \beta_m \cdot x_{im} + \varepsilon_i$$

bzw. als Modell mit Standardwerten:

$$\Rightarrow \hat{z}_{yi} = b_1 \cdot z_{i1} + b_2 \cdot z_{i2} + ... + b_m \cdot z_{im} + \varepsilon_i$$

Schätzung der Modellparameter

Zur Schätzung der unbekannten Modellparameter werden Datenstichproben aus a-priori definierten Grundgesamtheiten gezogen und die Koeffizienten derart bestimmt, dass die Summe der quadrierten Schätzfehler (Residualquadrate) minimal ist (Kleinst-Quadrate-Kriterium).

Fehler-/Residualquadratsumme: $\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \boldsymbol{e'e} = \text{SAQ} = \min$ (2)

Durch Einsetzen von $\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{b} + \boldsymbol{e}$ (ALM in Matrixschreibweise) in die obige Formel erhält man:

$$\begin{aligned} \text{SAQ} &= (\boldsymbol{y} - \hat{\mathbf{y}})'(\boldsymbol{y} - \hat{\mathbf{y}}) = (\boldsymbol{y} - \boldsymbol{Xb})'(\boldsymbol{y} - \boldsymbol{Xb}) = \\ &= \boldsymbol{y'y} + \boldsymbol{b'X'Xb} - 2\boldsymbol{b'X'y} \end{aligned} \quad (3)$$

Bildung der ersten Ableitung der obigen Beziehung nach $\boldsymbol{b'}$ und Nullsetzen ergibt:

$$\frac{\delta(SAQ)}{\delta \mathbf{b'}} = 2\boldsymbol{X'Xb} - 2\boldsymbol{X'y} = 0 \quad (4)$$

und führt zur Lösung:

$$\hat{\mathbf{b}} = (\boldsymbol{X'X})^{-1}\,\boldsymbol{X'y}$$

Modelltests

Die zentralen Prüffragen beziehen sich auf die statistische Absicherung des Einflusses einzelner oder aller Prädiktoren auf das Kriterium. In der Regel ist der erste Schritt die statistische Prüfung des vollen Modells. Ziel ist, festzustellen, ob die Modellprädiktoren überhaupt einen empirisch nachweisbaren Zusammenhang mit dem Kriterium haben. Die allgemeine Nullhypothese für die Prüfung des vollen Modells lautet:

- Die Prädiktoren x_1 bis x_m haben keinen Zusammenhang mit y

was der folgenden Behauptung gleichkommt:

- $\beta_1 = \beta_2 = ... = \beta_m = 0$

In weiteren Schritten werden dann in der Regel konkurrierende Modelle getestet, die den Einfluss einzelnen Prädiktoren behaupten. Ziel ist in der Regel die Bestimmung eines Modells, welches die Balance zwischen dem wissenschaftlichen Prinzip der Sparsamkeit und einer optimalen Prädiktion am besten repräsentiert. Die gängigen statistischen Programmpakete bieten dabei verschiedene Optionen zur schrittweise Prüfung von Alternativmodellen (bspw. „Stepwise Regression“ in SPSS). Zur Bestimmung der Modellgüte wird dabei das multiple Bestimmtheitsmaß als Indikator für die Prognosequalität der Prädiktoren heranzuziehen. Das multiple Bestimmtheitsmaß (multipler Determinationskoeffizient) R^2 setzt die die Varianz der Schätzwerte $Var(\hat{Y})$ - ins Verhältnis zur Varianz der empirischen Kriteriumswerte,

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)}$$

so dass die Nullhypothese bei der Prüfung des vollen Modells auch mit $H_0 = R^2 = 0$ formalisiert werden kann.

Anwendungsbeispiel: Studienerfolgsdeterminanten

Im Zentrum einer Untersuchung von Reiß und Moosbrugger (2004) stand die Frage, ob die retrospektive Einschätzung des Studiums von Absolventen des Diplomstudienganges Psychologie an der Universität Frankfurt einen Zusammenhang mit Studiendauer und den Studienerfolg aufweist. Es wurde versucht, sowohl aus den individuellen Unterschieden in den Studienbedingungen als auch aus den institutionellen Parametern des Studienganges jene Prädiktorvariablen zu identifizieren, welche einen Einfluss auf die Kriterien ‚Studiendauer’ und ‚Studienerfolg’ haben.

Die zur Vorhersage geeigneten Prädiktoren wurden hierzu inhaltliche in Blöcke eingeteilt. Ein erster Block mit dem Titel „individuelle Leistungsfähigkeit“ (1) beinhaltete u. a.

Prädiktoren wie „Abschlussnote im Vordiplom“. Der zweite Block bestand aus Variablen zum „universitären Engagement“ (2), der dritte fasste Prädiktoren zur „Organisation des Studium“ (3) zusammen, weitere Blöcke waren „Beratung und Betreuung durch die Universität“ (4), „Studienklima, Kontakte und Anbindung an die Universität“ (5) und schließlich „Organismusvariablen und individuelle Einstellungen“ (6).

Die einzelnen Prädiktoren des jeweiligen Blockes wurden nun unter Beachtung des größtmöglichen Inkrementes an determinierter Varianz schrittweise in eine multiple Regressionsanalyse aufgenommen. Dieses Vorgehen verhinderte Redundanzen, d.h. das Vorliegen einer Informationsüberschneidung der einzelnen Prädiktorvariablen.

Mit dem Kriterium Studiendauer korrelierten Variablen aus den Blöcken 1, 4, 5 und 6. Die folgende Tabelle zeigt ein Modell mit fünf Prädiktoren, welches zu einem multiplen Korrelationskoeffizienten von R = .60 führt. Mit einem Determinationskoeffizient von .36 klärt dieses Vorhersagemodell 36 Varianzprozente der Studiendauer auf.

Tab. 4:
Vorhersagemodell für das Kriterium Studiendauer (Anzahl der Studiensemester)

Modell	R	R^2
Abschlussnote Vordiplom	.41	.17
+ Inform. Berat. u. Betreu. durch Profs.	.47	.22
+ Ø-Note Studienberechtigung	.52	.27
+ EDV-Kenntnisse zu Studienbeginn	.57	.32
+ Kurze Studiendauer wichtig	.60	.36

Nach der gleichen Vorgehenslogik wurden aus den obigen sechs Variablenblöcken Prädiktoren in eine multiple Regressionsanalyse mit dem Kriterium Studienerfolg überführt.

Wie die folgende Tabelle zeigt, gelingt mit den dort aufgeführten sechs Prädiktoren eine erstaunlich präzise Prädiktion (multiples R = .75). Mit einem Determinationskoeffizienten von .56 klärt das Modell 56 % des Studienerfolgs auf.

Tab.5: Vorhersagemodell für das Kriterium Studienerfolg (Durchschnittsnote im Hauptdiplom)

Modell	R	R2
Abschlussnote Vordiplom	.55	.30
+ Alter bei Studienbeginn	.61	.37
+ Gewinn aus Selbststudium z. Vorb. auf Dipl.-Prüfung	.67	.45
+ Nutzung v. Vorlesungen z. Vorb. auf Dipl.-Prüfung	.70	.49
+ Engagement in Gremien	.73	.53
+ Inform. Berat. u. Betreu. durch WiMi	.75	.56

Literatur

Cohen (1968). Multiple Regression as a general data-analytic system. *Psychological Bulletin, 70,* 426-443.

Moosbrugger, H. (2003). *Lineare Modelle. Regressions- und Varianzanalysen* (3. Aufl). Bern: Huber.

Moosbrugger, H. & Frank, D. (1992). *Clusteranalytische Verfahren in der Persönlichkeitsforschung.* Bern: Huber.

Reiß, S. und Moosbrugger, H. (2004). Prädiktoren von Studiendauer und Studienerfolg: Ergebnisse einer Absolventenbefragung im Diplomstudiengang Psychologie der Goethe-Universität Frankfurt (WS 1995/96 –SS 2002), …

Steyer, R. (1992). *Theorie kausaler Regressionsmodelle*. Stuttgart: Gustav Fischer.

Erweiterte Regressionsmodelle

Wolfgang Rauch

Einführung

Regressionsanalyse wird grundsätzlich als eine Methode verstanden, bei der die Ausprägungen einer (oder mehrerer) abhängigen Variablen (AV; auch: Kriterium) auf die Ausprägungen in einer (oder mehrerer) unabhängiger Variablen (UV; auch: Prädiktoren) zurückgeführt werden. In den meisten Lehrbüchern der psychologischen Methodenlehre wird dabei unter dem Stichwort "Regressionsanalyse" im wesentlichen die lineare Regression im Rahmen des Allgemeinen Linearen Modells mit einer intervallskalierten AV und mehreren intervall- und/oder nominalskalierten UV und der Kleinstquadrateschätzung der Parameter behandelt. Dieses Modell soll im Folgenden als "ordinary least squares-Regression" (OLS-Regression) bezeichnet werden. Die Popularität dieses Modells ergibt sich zum einen aus seiner Einfachheit bei der Berechnung und Interpretation; zum anderen lässt sich auch zeigen, dass bei Gültigkeit der in diesem Modell nötigen Voraussetzungen die Kleinstquadrateschätzung optimale Schätzergebnisse liefert, nämlich unverzerrte Schätzungen mit der kleinstmöglichen Varianz (z.B. Werner, 1997). Allerdings sind die Voraussetzungen für die Kleinstquadrateschätzung im ALM relativ streng, so dass diese vorteilhafte Eigenschaft meist nicht gilt. Außerdem sind die Verwendungsmöglichkeiten des Modells eingeschränkt, da nicht alle Arten von Datenlagen damit analysiert werden können.

Mittlerweile gibt es eine Vielzahl von erweiterten Regressionsmodellen, die sich auch auf andere Variablenkonstellationen (z.B. ordinalskalierte abhängige Variablen) anwenden lassen oder deutlich bessere Anpassungen für nichtlineare Zusammenhänge erlauben. Darüber hinaus existieren auch annähernd modellfreie algorithmische Verfahren, die für regressionsanalytische Verfahren genutzt werden können. Einen Überblick über moderne regressionsanalytische Strategien gibt Harrell (2001).

In diesem Beitrag soll zunächst gezeigt werden, wie nicht-lineare Zusammenhänge zwischen Variablen im Rahmen der OLS-Regression modelliert werden können. Danach wird die Erweiterung auf das generalisierte lineare Modell vorgestellt und die Möglichkeiten nonparametrischer Regressionsverfahren gezeigt. Schließlich wird kurz auf algorithmische Modelle wie neuronale Netze und Klassifikationsbäume eingegangen.

Nicht-lineare Zusammenhänge in der OLS-Regression

Rückschau: ALM

Das allgemeine lineare Modell hat die Form

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i \qquad (1)$$

Die Ausprägung einer Person *i* in der *abhängigen Variable y* wird dabei als eine Linearkombination der Ausprägung von *i* in den *j unabhängigen Variablen x* dargestellt, d.h. als eine der mit den *Regressionsparametern* β_j gewichtete Summe der Ausprägungen x_i.

In Abbildung 1 ist exemplarisch ein lineares Regressionsmodell mit zwei unabhängigen Variablen dargestellt. Deutlich wird dabei der lineare Zusammenhang zwischen den zwei UV und der AV: Mit höheren Werten in x_1 und x_2 steigen die Ausprägungen von *y* linear an.

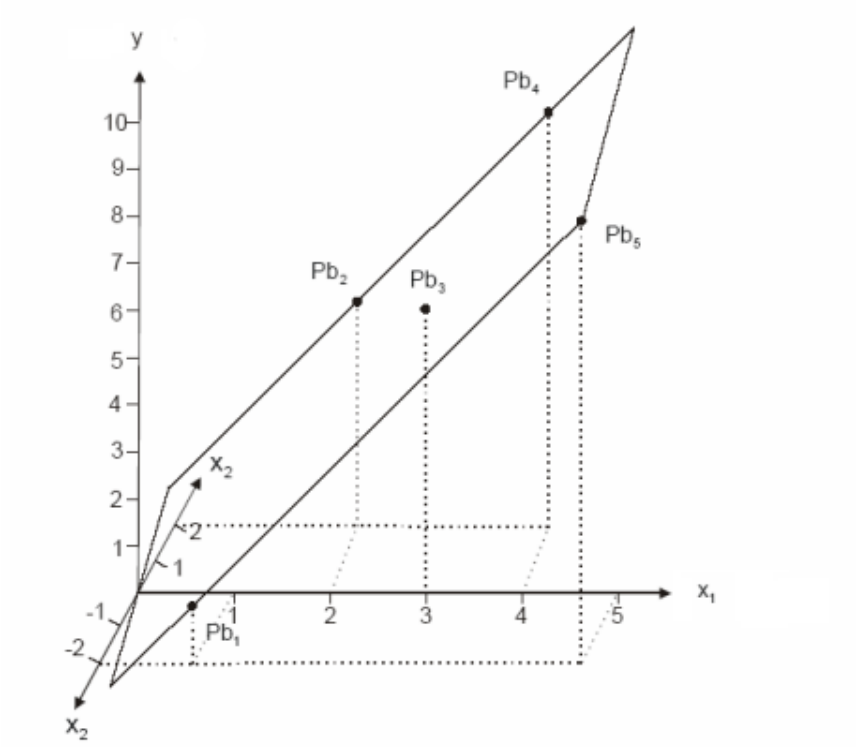


Abbildung 1: Lineares Regressionsmodell mit zwei unabhängigen Variablen x_1 und x_2 (entnommen aus Moosbrugger, 2003)

Kurvilineare Zusammenhänge

Oftmals ist ein einfacher linearer Zusammenhang zwischen den Variablen aber nicht angemessen. In Abbildung 2 ist ein kurvilinearer Zusammenhang zwischen der abhängigen Variable *y* und einer unabhängigen Variablen *x* dargestellt: Die Ausprägungen von *y* steigen umso stärker an, je höher *x* ausgeprägt ist. Die Anpassung nur eines linearen Zusammenhangs erbringt in diesem Beispiel eine deutlich geringere Varianzaufklärung als

die Anpassung des kurvilinearen Zusammenhangs. Zur Veranschaulichung sind in Abbildung 2 sowohl das lineare Regressionsmodell als auch das kurvilineare Modell eingezeichnet.

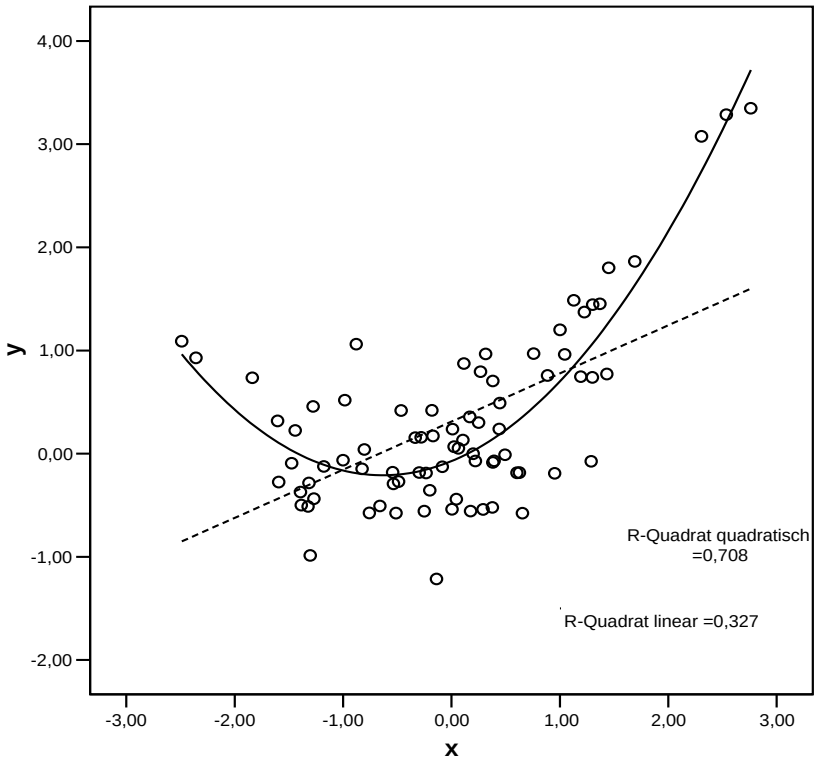


Abbildung 2: Lineares (gestrichelte Linie) und kurvilineares (durchgezogene Linie) Regressionsmodell für den Zusammenhang einer abhängigen Variablen y und einer unabhängigen Variablen x

Solche einfachen nicht-linearen Zusammenhänge lassen sich im Rahmen einer OLS-Regression auf unkompliziertem Wege modellieren: Dazu wird eine Transformation der unabhängigen Variablen vorgenommen, für die ein nicht-linearer Zusammenhang vermutet wird, und die transformierte Variable wird als eine weitere UV in das Modell mit aufgenommen. Das Modell für die gekrümmte Regressionslinie in Abbildung 2 etwa lautet:

$$y = -0,07 + 0,43x + 0,34x^2$$

Zur Darstellung kurvilinearer Zusammenhänge gibt es unterschiedliche Möglichkeiten, je nach Form des vermuteten Zusammenhangs. Besteht noch keine theoriegeleitete Annahme über die Form des Zusammenhangs, hilft die grafische Inspektion von Streudiagrammen (wie z.B. in Abbildung 2) bzw. bei multiplen Regressionen mit mehreren UV von Partial-Residualgrafiken, in denen die Residuen eines Regressionsmodells mit den übrigen UV gegenüber den Werten in der interessierenden UV abgetragen werden (vgl. Fox, 2000).

Moderatormodelle

In vielen psychologischen und sozialwissenschaftlichen Theorien werden auch andere Arten von nicht-linearen Zusammenhängen vermutet, nämlich die so genannten Moderator- oder Interaktionsmodelle. Von einer Interaktion zwischen zwei UV wird gesprochen, wenn eine Prädiktorvariable auf den Stufen einer anderen Prädiktorvariablen unterschiedliche Steigungen hinsichtlich der Kriteriumsvariablen aufweist (z.B. Moosbrugger, 2003). Ein inhaltliches Beispiel geben Hofmann, Rauch und Gawronski (in Vorb.): Als AV wurde die Menge an Schokolade gemessen, die eine Person in einer Geschmackstestaufgabe konsumierte. Die konsumierte Menge hing dabei von den selbstberichteten Diätstandards der Personen ab, aber für eine Gruppe von Personen, deren Selbstkontrollressourcen zuvor erschöpft worden waren, gab es keinen Zusammenhang der konsumierten Menge mit den Diätstandards. Anders ausgedrückt ist der Einfluss der Diätstandards unterschiedlich, je nachdem, wie erschöpft die Selbstkontrollressourcen einer Person sind. Diese Interaktion ist in Abbildung 3 veranschaulicht.

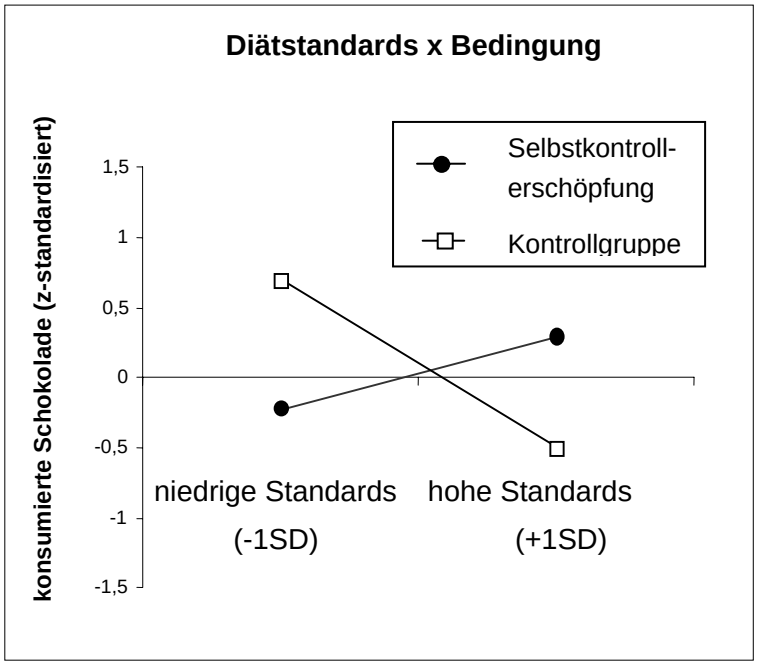


Abbildung 3: Interaktion zwischen Diätstandards und Versuchsbedingung bei der Vorhersage der Menge an konsumierter Schokolade. Entnommen aus Hofmann, Rauch und Gawronski (in Vorb.)

Auch Interaktionen zwischen UV lassen sich einfach in OLS-Regressionsmodelle einfügen. Dazu wird (nach Zentrierung der Variablen, vgl. Cohen, Cohen, West & Aiken, 2003) eine neue Variable aus dem Produkt der UV, für die eine Interaktion vermutet wird, gebildet und mit in das Regressionsmodell aufgenommen. Wie auch sonst ist es dabei nicht von Bedeutung, ob die beteiligten Prädiktoren intervall- oder nominalskaliert sind.

Zusammenfassung: Nicht-lineare Zusammenhänge im OLS-Modell

Zusammenfassend lässt sich festhalten, dass zumindest bestimmte Formen von nicht-linearen Zusammenhängen mittels des OLS-Regressionsmodells spezifizieren lassen. Anstelle der untransformierten Prädiktorvariablen werden dabei andere Variablen ins Modell aufgenommen, die eine Funktion der ursprünglichen Variablen darstellen, etwa $f(x) = x^2$ in ein Modell $\hat{y} = b_0 + b_1 f(x)$; alternativ können nicht-lineare Zusammenhänge modelliert werden, indem statt der AV eine Funktion der AV eingesetzt wird, beispielsweise die logarithmierte AV.

Ein Problem des OLS-Regressionsmodells ist es, dass die zugehörigen Annahmen sehr restriktiv sind und dass nicht alle möglichen Modelle spezifiziert werden können – Abhilfe schaffen etwa generalisierte lineare Modelle oder die nicht-parametrische Regression. Darüber hinaus gibt es Anwendungsfälle, in denen andere Modelle bessere Vorhersagen liefern, etwa Klassifikationsbäume oder neuronale Netze. Schließlich kommt es vor, dass die Form eines (nicht-linearen) Zusammenhangs nicht a priori bekannt ist; in solchen Fällen kommen ebenfalls nicht-parametrische Regressionsverfahren oder algorithmische Verfahren wie Klassifikationsbäume oder neuronale Netze zur Anwendung.

Generalisiertes Lineares Modell

Im OLS-Regressionsmodell können zwar auf einfachem Wege nicht-lineare Zusammenhänge spezifiziert werden, dennoch ist es in gewisser Hinsicht beschränkt: Als AV kann immer nur eine intervallskalierte Variable vorhergesagt werden und es muss eine Reihe von Voraussetzungen erfüllt sein, damit das Modell zu optimalen Schätzungen gelangt.

Betrachtet man beispielsweise den Fall einer dichotomen AV mit den Ausprägungen 0 und 1, so kann keine OLS-Schätzung mehr vorgenommen werden: Beispielsweise würde ein solches Modell zu vorhergesagten Werten führen, die außerhalb des möglichen Wertebereiches liegen oder die Größe der Residuen wäre abhängig von der Ausprägung der UV, was eine Verletzung ganz zentraler Annahmen des OLS-Modells darstellt. Doch auch für den Fall einer dichotomen AV gibt es ein Regressionsmodell, nämlich die so genannte *logistische Regression*. Bei der logistischen Regression wird als AV nicht direkt die Ausprägung (0 oder 1) vorhergesagt, sondern es wird das Verhältnis der Wahrscheinlichkeit, eine der Ausprägungen aufzuweisen ($\hat{p}_i(Y=1|X)$), und der Gegenwahrscheinlichkeit ($1-\hat{p}_i(Y=1|X)$) modelliert. In Gleichung 2 ist das Modell der logistischen Regression dargestellt.

$$\frac{\hat{p}_i(Y=1|X)}{1-\hat{p}_i(Y=1|X)} = e^{\left(\beta_0+\sum_{j=1}^{m}\beta_j x_{ij}\right)} \qquad (2)$$

In Abbildung 4 ist die Wahrscheinlichkeit $\hat{p}_i(Y=1|X)$ nach einem solchen logistischen Regressionsmodell dargestellt.

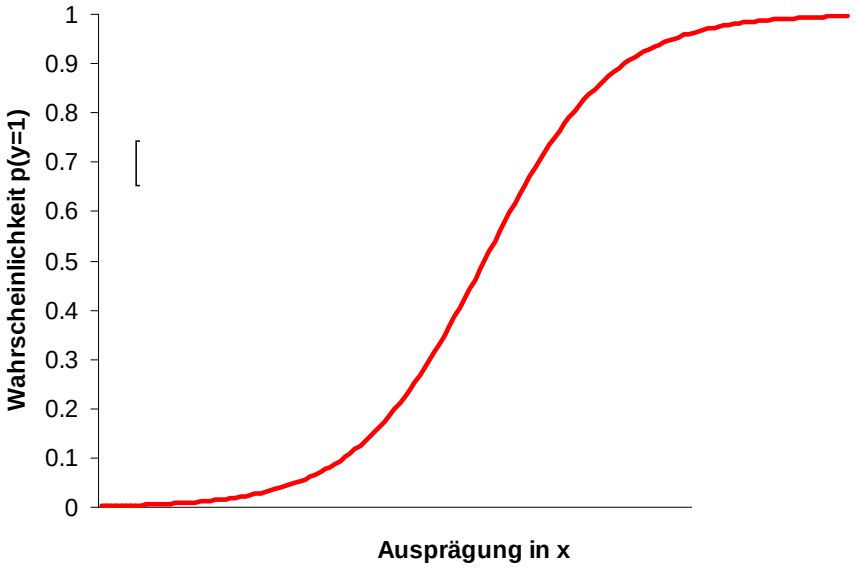


Abbildung 4: Veranschaulichung einer logistischen Regression: Die Wahrscheinlichkeit, dass die dichotome Variable y die Ausprägung 1 annimmt, steigt mit ansteigenden Werten in x

Das logistische Regressionsmodell gehört zu einer allgemeinen Klasse von Regressionsmodellen, in denen Funktionen der AV eingesetzt werden, um die Regressionsparameter zu schätzen, wie beispielsweise bei der logistischen Regression, die linear in den Parametern wird, indem der gesamte Ausdruck (2) logarithmiert wird:

$$h(y) = \ln\left(\frac{\hat{p}_i(Y=1|X)}{1-\hat{p}_i(Y=1|X)}\right) = \ln\left(e^{(\beta_0+\sum_{j=1}^{m}\beta_j x_{ij})}\right) = \beta_0 + \sum_{j=1}^{m}\beta_j x_{ij} \qquad (3)$$

In (3) bezeichnet *h(y)* die sog. Link-Funktion, die angewendet wird, damit das Modell linear in den Parametern wird. Wenn für $h(\cdot)$ eine beliebige Link-Funktion eingesetzt wird, spricht man vom *Generalisierten Linearen Modell* (GLM). Das GLM erweitert damit das ALM z.B. um die Möglichkeit der Regression auf nominal- oder ordinalskalierte AV; darüber hinaus können Modelle spezifiziert werden, für die die Voraussetzungen der OLS-Regression nicht gelten müssen. Eine allgemeine Form des generalisierten linearen Modells ist in (4) dargestellt; darin bezeichnet $C(Y|X)$ eine Verteilung von Y als eine Funktion von X mit einer Link-Funktion $h(\cdot)$.

$$h(C(Y|X)) = X\beta \qquad (4)$$

Im generalisierten linearen Modell kommen auch andere Schätzverfahren zur Anwendung als die Kleinstquadrateschätzung, etwa die Maximum-Likelihood-Schätzung in der logistischen Regression.

Nicht-parametrische Regression

Nicht-parametrische Regressionsverfahren (z.B. Fox, 2000) erweitern auch noch das Modell der generalisierten linearen Regression (4). Während im GLM noch a priori eine bestimmte Form der Regressionsfunktion a priori angenommen wird bzw. eine bestimmte Verteilung der AV unterstellt wird (daher die Bezeichnung parametrisch), werden bei nicht parametrischen Verfahren keine solchen a priori Annahmen getroffen, sondern die Regressionsfunktion direkt aus den Daten geschätzt. Grundsätzlich haben solche Modelle eine solche Form, dass nicht für alle Ausprägungen der UV ein und dieselbe Regressionsfunktion angepasst wird, sondern dass für bestimmte Ausprägungen $X = x$ unterschiedliche Funktionen angepasst werden, wobei unterschiedliche Anpassungsmethoden vorliegen:

$$h(C(Y \mid X = x)) = X\beta \tag{5}$$

Einen Überblick über non-parametrische Regressionsverfahren gibt Fox (2000).

Algorithmische Verfahren

Neuronale Netze und Regression

Neuronale Netze, eine Entwicklung aus der Forschung über künstliche Intelligenz, werden ebenfalls häufig für regressionsanalytische Probleme eingesetzt (Warner & Misra, 1996). Ein (künstliches) neuronales Netz ist ein Versuch, die Informationsverarbeitung in menschlichen und tierischen Nervenzellen und deren Verbindungen untereinander nachzubilden. Dabei werden mehrere künstliche "Neuronen" programmiert, die miteinander verknüpft sind und einander Informationen weitergeben. Typischerweise wird ein gerichteter Informationsfluss modelliert dergestalt, dass Eingangsneuronen Input erhalten und verarbeiten und diesen Output an eine nächste Ebene weitergeben, wobei die Zahl der Neuronen innerhalb einer Ebene sowie die Zahl der Ebenen beliebig festgelegt werden können. Zentral ist dabei die Verarbeitung von Input zu Output: Der Input in ein Neuron *i* ist dabei eine der mit den Gewichten w_{ij} gewichtete Summe der Outputs der Neuronen *j*, die mit *i* verknüpft sind:

$$input_i = \sum_j w_{ij} \cdot output_j + \mu_i \tag{6}$$

Auf dieser Ebene wird die Analogie zu einem einfachen Regressionsmodell deutlich. Der Input wird dann meist noch weiter verrechnet, um den Output des Neurons zu bestimmen, der wiederum an andere Neurone weitergegeben wird.

Das grundlegende Problem in neuronalen Netzen ist nun, die Gewichte w_{ij} so zu bestimmen, dass ein bestimmtes Ziel erreicht wird. In einem Regressionsproblem etwa könnte ein neuronales Netz dazu benutzt werden, die Summe der quadrierten Abweichungen zwischen der AV und dem vom Netz vorhergesagten Wert in der AV zu minimieren. Dazu wird ein algorithmisches Vorgehen gewählt:

(a) Die Ausprägungen von UV und AV werden in das Netz eingegeben und eine Vorhersage aufgrund der bestehenden Gewichte getroffen

(b) diese Vorhersage wird mit der tatsächlichen Ausprägung verglichen

(c) es wird geprüft, ob eine Veränderung der Gewichte die Vorhersage verbessert

(d) wenn sich die Vorhersage verbessert, werden die Schritte (a) bis (c), wiederholt, wenn nicht, wird das Modell nicht mehr verändert

In Analogie zu einem nicht-parametrischen Regressionsmodell wird die Vorhersage in einem neuronalen Netz also nicht nach einer a priori-Spezifikation getroffen, sondern anhand der vorliegenden Daten angepasst. Die Verknüpfung der einzelnen Neuronen untereinander kann betrachtet werden wie das Problem, eine jeweils passende Funktion der Prädiktorvariablen zu finden, um die AV möglichst genau vorherzusagen.

Neuronale Netze sind sehr flexibel in Bezug auf die Anzahl der Neuronen, die Anzahl der Ebenen, die Verrechnung des Inputs zu einzelnen Neuronen und die Art der Anpassung der Gewichte. Während sie auf der einen Seite unter Umständen eine sehr genaue Vorhersage erlauben, sind die Ergebnisse auf der anderen Seite meist schwer interpretierbar.

Klassifikations- und Regressionsbäume (CART)

Wie neuronale Netze nutzt auch das CART-Verfahren ein algorithmisches Vorgehen für die Vorhersage (Harrell, 2001). Das Vorgehen ist in den folgenden Schritten zusammengefasst:

(a) Finde den Prädiktor, so dass die bestmögliche binäre Trennung auf diesem Prädiktor den Wert in einem statistischen Kriterium (z.B. Kleinstquadratekriterium) erreicht, der besser ist als alle anderen Trennungen auf allen anderen Prädiktoren, wobei sich das statistische Kriterium auf einer Zusammenfassung der Beobachtungen (z.B. dem Mittelwert) beruht

(b) In jeder so geformten Untermenge finde den besten Prädiktor mit der bestmöglichen Trennung wie in (a)

(c) Wiederhole (b) so lange, bis die Untermengen eine bestimmte Größe unterschreiten

Um eine Überanpassung des Modells zu vermeiden, werden die Schritte (a) bis (c) jeweils an einem Training-Anteil der Daten durchgeführt und an einem Test-Anteil validiert, so lange, bis der bestmögliche Baum gefunden wurde.

Hypothesenprüfung und Vorhersage bei regressionsanalytischen Fragestellungen

Beim Einsatz eines nicht-parametrischen Regressionsmodells oder eines algorithmischen Verfahrens ist das wesentliche Ziel eine möglichst gute Anpassung des Modells an die Daten im Unterschied zu einer theoriegeleiteten Prüfung eines a priori festgelegten Zusammenhangs. Solche mehr oder weniger modellfreien Verfahren haben gegenüber einem theoriegeleiteten Vorgehen den Vorteil, dass sie meist eine deutlich genauere Vorhersage der AV erlauben als a priori spezifizierte parametrische Modelle. Auf der anderen Seite haben nicht-parametrische oder modellfreie Zugänge den Nachteil, dass sie meist nur schwer interpretierbar sind und Schlüsse auf Kausalzusammenhänge der beteiligten Variablen deutlich erschwert werden.

Vor der Anwendung regressionsanalytischer Verfahren muss also das Ziel der Analyse eindeutig geklärt werden: Für den Fall, dass Hypothesen über vorher spezifizierte Zusammenhänge geprüft werden sollen, ist ein parametrisches Regressionsmodell anzuwenden, da die Interpretation dort sehr einfach ist und die Eigenschaften der Hypothesentests bekannt und vielfach erprobt sind. Besteht das Untersuchungsziel aber in einer möglichst guten Vorhersage des Kriteriums unabhängig davon, welche Prädiktoren in welcher Form zur Vorhersage beitragen, sollte ein nicht-parametrisches Modell oder ein modellfreies Vorgehen gewählt werden, die meist eine deutlich größere Genauigkeit der Vorhersage erlauben (vgl. z.B. Breiman, 2001).

Literatur

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statistical Science, 16*, 199-231.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3rd Ed.). Mahwah, N.J.: Lawrence Erlbaum.

Fox, J. (2000). *Nonparametric Simple Regression: Smoothing Scatterplots.* (Sage University Papers Series Quantitative Applications in the Social Sciences, series no. 07-130). Thousand Oaks, CA: Sage.

Harrell, Jr., F. E. (2001). *Regression Modeling Strategies.* New York: Springer.

Hofmann, W., Rauch, W. & Gawronski, B. (in Vorb.). And Deplete us not into Temptation: Implicit Attitudes, Dietary Restraint, and Self-Regulatory Resources as Determinants of Eating Behavior. *Unveröffentlichtes Manuskript.*

Moosbrugger, H. (2003). *Lineare Modelle. Regressions- und Varianzanalysen*. Bern: Huber.

Warner, B. & Misra, M. (1996). Understanding Neural Networks as Statistical Tools. *The American Statistician, 50,* 284-293.

Werner, J. (1997). Lineare Statistik: Das Allgemeine Lineare Modell. Weinheim: Beltz, PVU.

Smoothing und non-parametrische Regression

Christine Berude und Samantha Wasser

Was ist Smoothing und non-parametrische Regression?

Hinter den Begriffen Smoothing und non-parametrische Regression verbergen sich Verfahren, die eine Erweiterung des linearen Regressionsmodells darstellen und sich dabei dem Problem stellen, wie man sich Beziehungen zwischen Variablen nährt, wenn man noch keine klare Vorstellung über deren Form hat oder eine solche erst einmal nicht formulieren möchte.

Der zentrale Unterschied zwischen parametrischer Regression und non-parametrischer Regression besteht also in den Vorannahmen, die man explizit formuliert. Um eine parametrische Regressionsanalyse durchführen zu können ist es notwendig, sich *vorher* Gedanken über die Art der Beziehung zwischen den Variablen zu machen. Dementsprechend wird ein Modell formuliert, dass dann mit Informationen aus den Daten gefüllt wird.

Dem gegenüber ist es für die Anwendung eines Verfahrens zur non-parametrischen Regression nicht notwendig, sich vor der Analyse auf ein bestimmtes Modell festzulegen. Vielmehr dient das Verfahren der Exploration der interessierenden Beziehungen.

Um das eben Gesagte zu verdeutlichen, wird im Folgenden die grundlegende Vorgehensweise beider Ansätze kurz an einem Beispiel beschrieben. Ausgangspunkt ist jeweils das Streudiagramm zwischen einer Kriteriumsvariablen (Studienerfolg) und einer Prädiktorvariablen (Kreativität).

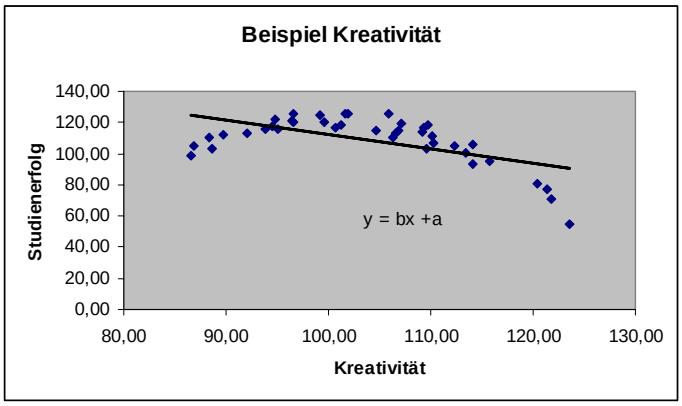


Abbildung 1: Parametrische Regression: Regressionskurve für die Beziehung zwischen Kreativität und Studienerfolg unter Voraussetzung eines Regressionsmodells vom Typ y=bx+a.

Abbildung 1 zeigt die Regressionsgerade zwischen Kreativität und Studienerfolg. Im Vorhinein wurde also festgelegt, dass diese Gerade der Form y=bx+a folgen soll, z. B. weil diese Art der Beziehung aus bestimmten Hypothesen hervorging.

Wenn man allerdings die Form des Punkteschwarms betrachtet, wird deutlich, dass eine Regressionsgerade den Punkteschwarm nicht sehr zutreffend beschreibt. Besser wäre es, eine andere Form des Zusammenhangs zu unterstellen. Diese „andere Form des Zusammenhangs“ könnte man nun in einem anderen Modell formulieren, von dem man meint, dass es die Daten besser trifft (ein Modell, das einen quadratischen Term von x beinhaltet wäre naheliegend), oder man könnte sich erst einmal an die Form des Zusammenhangs herantasten, ohne weitere Vorüberlegungen anzustellen. Diese zweite Möglichkeit beschreibt das Vorgehen bei der non-parametrischen Regression, deren Resultat, bezogen auf unser Beispiel, so aussehen könnte wie in Abbildung 2.

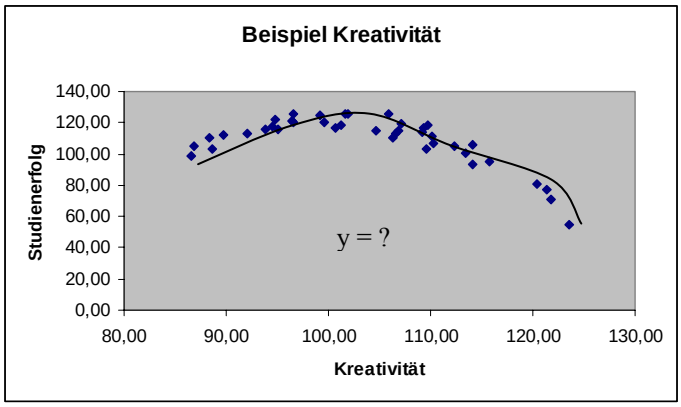


Abbildung 2: Non-parametrische Regression: Die Kurve, die zu den Daten passt, wird gesucht. Es existiert keine (explizit formulierte) Vorstellung über den Zusammenhang der Variablen.

Der Ausdruck y=? in Abbildung zwei bezieht sich noch einmal auf das fehlende Modell, welches einen Zusammenhang beschreibt. Ein Anwendungsfall der non-parametrischen Regression kann neben der grundlegenden Aufdeckung von Zusammenhängen auch die

Absicherung eines vermuteten Zusammenhangsmusters sein, wenn man, wie in eben gezeigtem Beispiel, bereits eine Ahnung hat, welche Form der Regressionskurve nahe liegend wäre. Prinzipiell entspricht das Absichern eines vermuteten Zusammenhangs per non-parametrischer Regression im Vergleich zur Formulierung eines neuen Modells einem vorsichtigeren und flexibleren Vorgehen, da auch unvermutete Zusammenhänge deutlich werden können.

Da das Vorgehen bei der non-parametrischen Regression einer Glättung der Punktewolke im Streudiagramm (=scatterplot) entspricht (indem, wie auch bei der parametrischen Regression, $\hat{y}$ -Werte entlang einer glatten Kurve geschätzt werden), wird die non-parametrische Regression oft auch als „scatterplot smoothing“ bezeichnet.

Im Folgenden wird das Vorgehen bei der non-parametrischen *Einfach*regression (es gibt also nur eine Prädiktorvariable) beschrieben, welches prinzipiell jedoch auf die multiple non-parametrische Regression (also mit mehreren Prädiktoren) verallgemeinert werden kann.

Zunächst werden jedoch die Daten, die zwecks der Vorstellung der verschiedenen Verfahren beispielhaft analysiert werden, kurz inhaltlich vorgestellt.

Einführung der verwendeten Beispiele

Die hier dargestellten Verfahren sollen anhand von zwei Datenbeispielen erläutert werden. Zum einen handelt es sich dabei um die (fiktive!) Beziehung zwischen Kreativität und Studienerfolg, auf die im letzten Abschnitt bereits zurückgegriffen wurde. Abbildung 3 zeigt das Streudiagramm der Daten ohne Regressionskurve, wie es typischerweise der Ausgangspunkt non-parametrischer Regression ist.

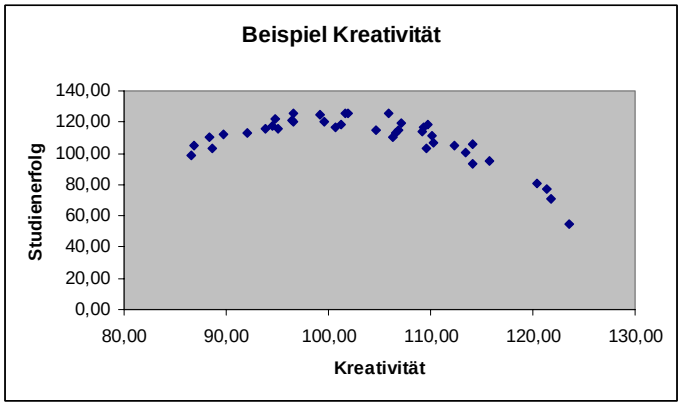


Abbildung 3: Streudiagramm der Variablen Kreativität und Studienerfolg.

Es wird deutlich, dass der Studienerfolg bis zu einem gewissen Punkt der Kreativität ansteigt, danach jedoch wieder abfällt. Zu den tatsächlichen Zahlenwerten ist zu sagen, dass diese für die Erläuterung der Verfahren unerheblich sind, wobei allerdings deutlich gemacht werden muss, dass hohe Werte jeweils für eine hohe Ausprägung des betreffenden

Merkmals stehen (kreativ bzw. erfolgreich). Von einem einfachen linearen Zusammenhang („je kreativer, desto erfolgreicher bzw. auch umgekehrt“) kann also nicht ausgegangen werden.

Das zweite (fiktive) Beispiel beschäftigt sich mit der Beziehung zwischen dem eines Studierenden zur Verfügung stehendem Geld (in Euro pro Jahr; ohne eigene Erwerbstätigkeit) und dem Studienerfolg. Letzterer ist hier operationalisiert als Erfolgsmessung nach einer bestimmten Studiendauer (z. B. vergleichbar mit der Note im Vordiplom im Studiengang Betriebswirtschaftslehre; das Vordiplom muss bis zu einer gewissen Semesteranzahl erworben werden).

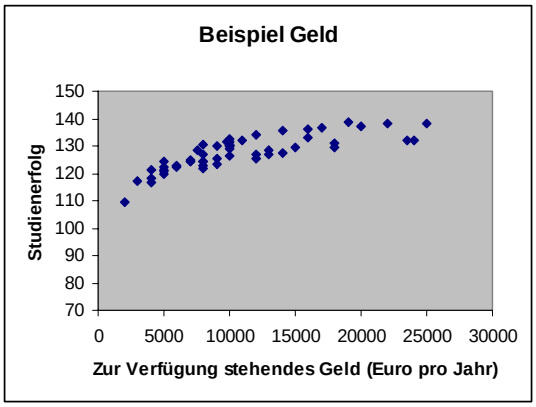


Abbildung 4: Streudiagramm der Variablen Geld und Studienerfolg.

Wie aus Abbildung 4 ersichtlich wird, sind Studierende mit relativ wenig zur Verfügung stehendem Geld nach dem eben beschriebenen Kriterium weniger erfolgreich als solche mit mehr Geld, was daran liegen könnte, dass sie mehr Zeit mit der Bestreitung des Lebensunterhalts verbringen müssen als ihre Kommilitonen und deswegen weniger Zeit zum Lernen haben. Allerdings werden die Unterschiede bezüglich des Studienerfolgs mit steigendem Auskommen immer geringer.

Binning und Lokale Mittelwertberechnung

Die beiden Methoden Binning und Lokale Mittelwertberechnung verfolgen dasselbe Ziel. Es geht darum, zu untersuchen, wie sich die Variable y verändert in Abhängigkeit der Variablen x. Wenn die Prädiktorvariable x diskret ist, also nur bestimmte Werte annehmen kann, ist es sehr einfach, den Zusammenhang zwischen den Variablen zu erkennen. Hier helfen uns die Mittelwerte von y über einen x Wert weiter. Die bedingten Mittelwerte der Population können direkt gemessen werden, wenn Daten einer ganzen Population vorhanden sind. Bei einer großen Stichprobe sind die gemessenen Mittelwerte sehr nahe den tatsächlichen Mittelwerten der Population. Man könnte auch den Median und die Quartile der Verteilung untersuchen, statt die Mittelwerte. Dies wäre bei einer verzerrten Verteilung auch sinnvoller.

Binning

Das Verfahren des Binnings, oder das Kästchen-Bildens, wird in Fällen angewandt, in denen die Prädiktorvariable x stetig ist, also unendlich viele Werte einnehmen kann. Hier entstehen Probleme bei der Errechnung der Mittelwerte, da sie zum Teil auf nur wenigen Beobachtungen oder einer einzigen Beobachtung errechnet wird. Es gibt auch Stellen, an denen keine Beobachtung vorhanden ist. Nehmen wir einmal an, die Prädiktorvariable sei, wie in unserem Beispiel, das zur Verfügung stehende Geld von Studenten. Anstatt dieses aber (wie üblich) gerundet auszudrücken, wird es auf den Cent genau angegeben. Sogar bei einer großen Stichprobe gäbe es hier wenige Personen mit genau dem gleichen Einkommen. Es gäbe auch viele Werte, die gar nicht besetzt wären. Wenn man nun die Mittelwerte für jedes x errechnen wollte, wären sie sehr variabel und eine schlechte Schätzung der Population.

Aus diesem Grund kann man die Stichprobe entlang des Range von x in mehrere Klassen bzw. Bins einteilen, wobei x_1, x_2, …, x_b den x-Wert in der Mitte des Bins repräsentiert. Bei einer großen Stichprobe enthält jedes Bin viele Daten, so sind die Stichproben-Mittelwerte für jedes Bin stabil. Dieser Bin-Mittelwert ist gut geeignet zur Schätzung der $\hat{y}$-Werte der x-Werte in der Kategorienmitte, die dann zur Regressionskurve verbunden werden. Abbildung 5 zeigt dieses Vorgehen für das Beispiel Studenten und zur Verfügung stehendes Geld.

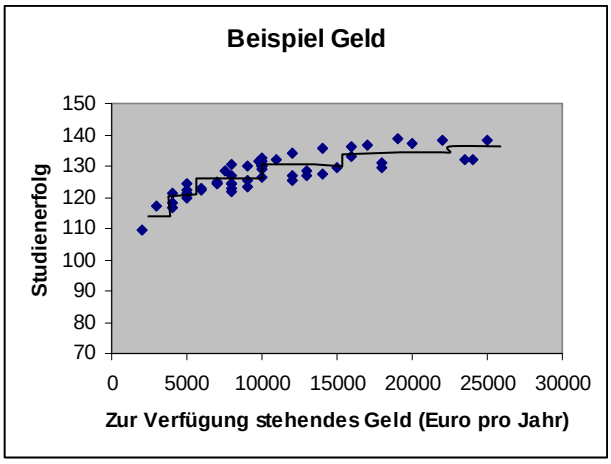


Abbildung 5: Die $\hat{y}$ -Werte wurden aus den Bin-Mittelwerten geschätzt und dann zur Regressionskurve verbunden.

Wenn genug Daten vorhanden sind, gehen mit dem Binning keine Kosten einher, aber in kleineren Stichproben ist es eher unpraktisch den Range von x in mehrere schmale Bins zu unterteilen. So wären in jedem Bin nur wenige Beobachtungen enthalten, was zu einer Instabilität der Stichproben-Mittelwerte der Bins führen würde. Um stabile Mittelwerte zu erhalten, muss man eine geringe Anzahl breiter Bins verwenden. In diesem Fall entsteht eine ungenaue Schätzung der Regressionsfunktion für die Population.

Es gibt zwei verschiedene Methoden, die Spanne von x zu unterteilen.

1. Man teilt den Range von x in gleich breite Bins.
2. Man teilt den Range von x in Bins, die ungefähr die gleiche Menge von Beobachtungen enthalten.

Die erste Methode ist nur sinnvoll, wenn x gleich verteilt ist. Wenn x nicht gleich verteilt ist, wären manche Bins vielleicht sogar leer. Aus diesem Grund wird meist die zweite Methode des Binning gewählt, nämlich unterschiedlich große Bins einzusetzen, die alle die gleiche Menge Daten beinhalten.

Schmale Bins werden bevorzugt, um den Bias zu verringern, welcher die Differenz zwischen dem Erwartungswert von y und dem geschätzten $\hat{y}$-Wert beschreibt. Der Bias ist in schmalen Bins klein und in großen Bins groß, weswegen schmale Bins bevorzugt werden. Hier stößt man jedoch auf Probleme bei der Schätzung des Bin-Mittelwerts, denn in schmaleren Bins sind oft zu wenig Daten enthalten. So entstehen hoch variable Bin-Mittelwerte. Man will natürlich versuchen sowohl den Bias als auch die Varianz gering zu halten, doch diese beiden Ziele arbeiten gegen einander. Breite Bins produzieren eine geringe Varianz, dafür aber einen hohen Bias; schmale Bins führen zu großer Varianz, aber der Bias bleibt gering. Bei einer genügend großen Stichprobe verschwinden diese Probleme: Man wählt schmale Bins und diese enthalten auch genügend Daten. Die non-parametrische Regression stößt immer wieder an das Problem der Varianz vs. dem Bias. Wenn die Regressionsfunktion der Population relativ weich verläuft, versucht man das Problem zu umgehen, indem man die Breite der Bins bis auf 0 verringert, während die Stichprobengröße n wächst. Die Breite der Bins muss aber langsam verringert werden, so dass die Anzahl der Beobachtungen innerhalb des Bins steigt. Unter diesen Umständen wird $bias[\hat{f}(x)] \rightarrow 0$ und $V[\hat{f}(x)] \rightarrow$ `0 während n` $\rightarrow \infty$.

Das Binning von stetigen Prädiktorvariablen wird oft bei der Analyse von großen Datensätzen durchgeführt, manchmal sogar schon bei der Datensammlung, wenn das Einkommen in Klassenintervallen angegeben werden soll, z.B. €0-4999, €5000-9999, €10000-14999, usw.

Lokale Mittelwertberechnung

Die Methode der lokalen Mittelwertberechnung geht erstmal davon aus, dass bei einer weich verlaufenden Regressionsfunktion die Beobachtungen mit x-Werten in der Nähe eines Blickpunktes x_0, Informationen liefern über $f(x_0)$. Die lokale Mittelwertberechnung ist sehr ähnlich dem Binning. Anstatt dass die Daten in nichtüberlappende Bins geteilt werden, wird ein Bin, hier allerdings Fenster genannt, kontinuierlich über die Daten bewegt und die Mittelwerte der Beobachtungen, die sich gerade im Fenster befinden, werden errechnet.

Es ist natürlich nicht umsetzbar, die Regressionsfunktion an einer unendlich großen Anzahl von Stellen zu schätzen. Man kann aber $\hat{f}(x)$ an einer großen Anzahl von Blickpunkten errechnen, normalerweise gleich verteilt über den Range der beobachteten x-Werte, oder an den Stellen der Beobachtungen $x_{(1)}, x_{(2)},\ldots, x_{(n)}$.

Wie beim Binning verwendet man nun ein Fenster mit festgelegter Breite, dessen Mittelpunkt auf den Wert des Blickpunktes x_0 zentriert ist, oder man passt die Breite des Fensters an, sodass es immer die gleiche Anzahl von Beobachtungen m beinhaltet. Diese sind die *m nächsten Nachbarn* des Blickpunktes.

Probleme können auftreten in der Nähe der Extremen von x. Diese werden boundary bias, oder Randbias genannt, da die Werte am Rand der Regressionskurve, also in der Nähe von $x_{(1)}$ und $x_{(n)}$ eine künstliche Abflachung der Kurve produzieren. Eine Lösung dieses Problems wäre, zu verlangen, dass gleich viele Nachbarn rechts und links von dem Blickpunkt im Fenster sein müssen. Solche symmetrischen Nachbarschaften beinhalten aber eine immer geringer werdende Anzahl von Beobachtungen, je weiter man sich dem Rand der Daten nähert, sodass oft die einzige Beobachtung in der Nachbarschaft von $x_{(1)}$ selbst $x_{(1)}$ ist. Eine bessere Lösung für Randbias wird in der Diskussion über lokale Regression aufgegriffen.

Abbildung 6 zeigt die Regressionskurve für unser Beispiel, die über die lokale Mittelwertberechnung erstellt worden ist.

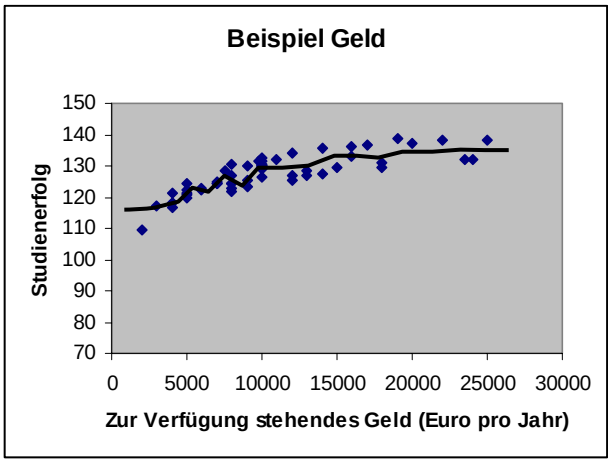


Abbildung 6: Non-parametrische Regression mit dem Lokalen-Mittelwerte-Verfahren.

Angepasste Werte (Schätzwerte $\hat{y}$) werden für jeden Blickpunkt x errechnet und dann miteinander verbunden. Zusätzlich zu dem Abflachen der Kurve rechts und links, produziert die lokale Mittelwertberechnung eine grob verlaufende Regressionskurve, da $\hat{f}(x)$ kleine Sprünge macht mit dem Herein- und Heraustreten der Beobachtungen in das Fenster. Der Kernel Schätzer, der im Anschluss beschrieben wird, produziert eine viel weicher verlaufende Regressionskurve.

Die Kernel Schätzung

Dieses Verfahren, auch lokale gewichtete Mittelwertberechnung genannt, ist eine Erweiterung der lokalen Mittelwertberechnung. Die Methode zielt darauf ab, bei der Schätzung von $f(x_0)$, eine größere Gewichtung denjenigen Beobachtungen zuzuteilen, die näher an dem Blickpunkt x_0 liegen, und die entfernt liegenden Beobachtungen weniger zu gewichten. Lass $z_i=(x_i-x_0)/h$ die skalierte, mit Vorzeichen behaftete Distanz zwischen dem x-Wert der i-ten Beobachtung und dem Blickpunkt x_0 bezeichnen. Der Skalierungsfaktor h, der als Bandbreite des Kernel Schätzers bezeichnet wird, spielt eine ähnliche Rolle wie die Fensterbreite bei der lokalen Mittelwertberechnung.

Zunächst benötigt man eine Kernel Funktion K(z), die das größte Gewicht an den Beobachtungen in der Nähe von x_0 anheftet und dann symmetrisch und weich verlaufend abfällt, je größer |z| wird. Solange die Funktion diese Voraussetzungen erfüllt, ist es relativ unkritisch welche Kernel Funktion man wählt. Es gibt nämlich verschiedene Funktionen, die man einsetzen könnte. Bei allen muss man zuerst die Gewichte $w_i=K[(x_i-x_0)/h]$ errechnen, um dann einen angepassten Wert an der Stelle x_0 zu errechnen durch gewichtete lokale Mittelwertberechnung der Werte von y.

Hierzu gibt es, wie bereits erwähnt, mehrere Methoden. Der Gauss'sche, oder normale, Kernel und der Trikubische Kernel sind Methoden, die sehr oft verwendet werden.

Der normale Kernel folgt der Standardnormalverteilung, bei der die Bandbreite h die Standardabweichung einer Normalverteilung ist, deren Mittelwert bei x_0 liegt. Beobachtungen, die eine Distanz größer als 2h vom Blickpunkt haben, sind somit mehr als zwei Standardabweichungen vom Mittelwert entfernt und erhalten eine Gewichtung nahe 0, weil dort die Funktionswerte klein sind.

Der Trikubische Kernel setzt die Bandbreite h mit der Breite des halben Fensters gleich, dessen Mittelpunkt an dem Blickpunkt x_0 zentriert ist. Beobachtungen, die außerhalb des Fensters liegen, erhalten die Gewichtung 0.

Eine weitere Methode ist der rechteckige Kernel, der alle Beobachtungen in einem Fenster, das die halbe Breite h hat, gleich gewichtet. So entsteht eine ungewichtete lokale Mittelwertberechnung, wie bereits beschrieben.

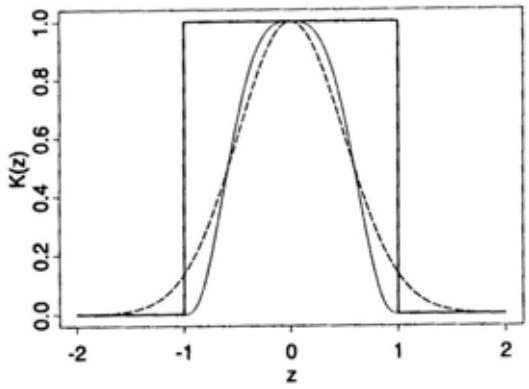


Abbildung 1: Trikubische (helle durchgezogene Linie), normale (gebrochene Linie) und ungewichtete (dunkle durchgezogene Linie) Kernel Funktionen aus Non Parametric Regression – Scatterplot Smoothing (Fox,2000a, S.18)

Bisher wurde davon ausgegangen, dass h eine feste Breite hat, aber der Kernel Schätzer kann leicht adaptiert werden, um eine bestimmte Anzahl nächster Nachbarn zu enthalten. Die Anpassung ist am einfachsten bei den Kernel Funktionen, die rechts und links zu 0 verlaufen, wie der Trikubische Kernel. Hier wird h(x) angeglichen, um eine feste Anzahl von Beobachtungen m im Fenster zu enthalten. Der Bruch m/n ist die Spanne (span) des Kernel Smoothers. Die Kernel Funktion wird Smoother genannt, weil er dazu führt, die Regressionskurve weicher verlaufen zu lassen. Wie bei der lokalen Mittelwertberechnung wird der Kernel Schätzer für eine Anzahl von Werten errechnet, die gleich verteilt sind über den Range von x, oder an den Beobachtungspunkten $x_{(i)}$. Die Kernel Schätzer (oder $\hat{y}$-Werte) werden dann zur Regressionskurve verbunden. In Abbildung 7 a-d ist das schrittweise Vorgehen bei der Kernel-Schätzung dargestellt.

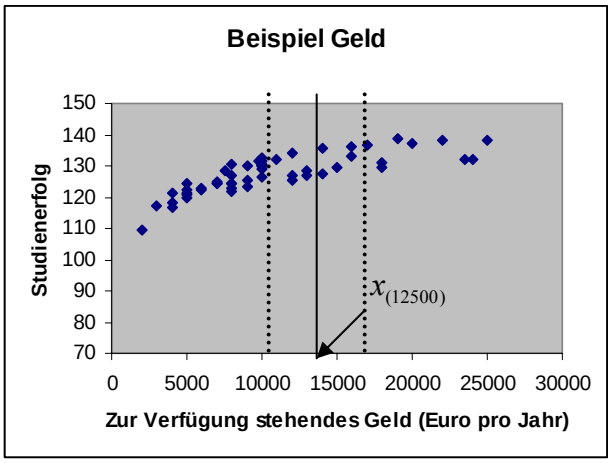


Abbildung 7a): Das Fenster schließt die m=15 nächsten Nachbarn ein.

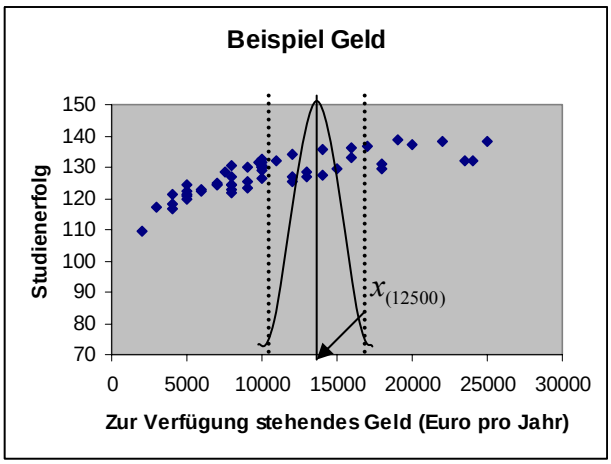


Abbildung7b): Die Trikubische-Kernel-Gewichtsfunktion.

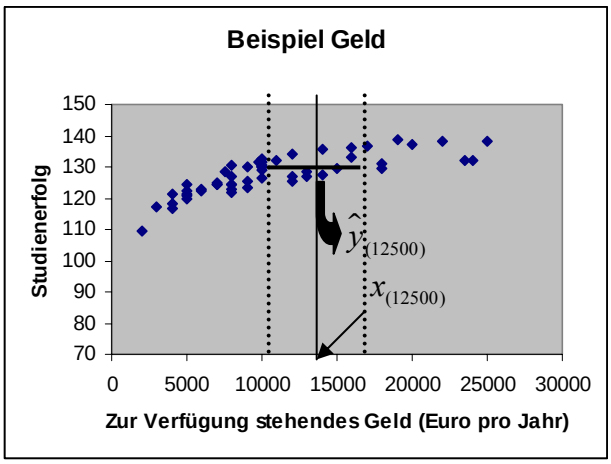


Abbildung 7c): Der $\hat{y}$ -Wert wird geschätzt aus dem mit den Kernel-Gewichten gewichteten Mittelwert des Fensters.

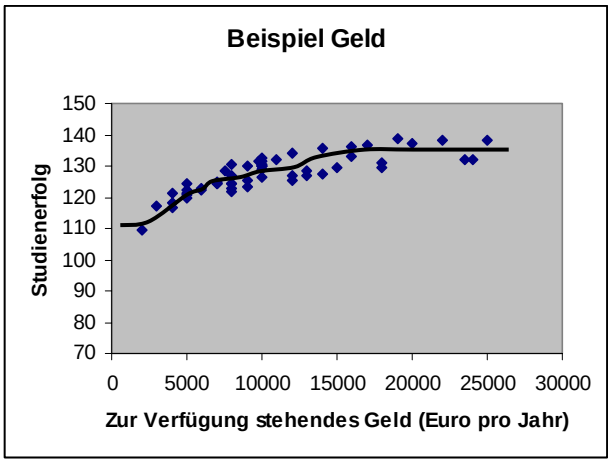


Abbildung 7d): Die Verbindung aller $\hat{y}$ -Werte ergibt die Regressionskurve.

Die Veränderung der Bandbreite des Kernel Schätzers ändert die Smoothness der geschätzen Regressionsfunktion. Größere Bandbreiten produzieren weicher verlaufende Kurven.

Lokale polynomiale Regression

Die lokale polynomiale Regression ist in der Lage, einige der Nachteile, die die Kernel-Schätzung mit sich bringt, auszugleichen und stellt einen generell adäquaten Ansatz zur non-parametrischen Regression dar (Fox, 2000a).

Das Ziel der lokalen polynomialen Regression besteht, wie auch schon beim Binning- und Lokalen Mittelwerte-Verfahren und der Kernel-Schätzung darin, für jeden interessierenden x-Wert x_0 einen $\hat{y}$-Wert zu schätzen. Alle diese geschätzten $\hat{y}$-Werte lassen sich dann, wie gehabt, zu einer Regressionskurve verbinden.

Wo jedoch bei der Kernel-Schätzung an den Enden der Kurve ein Abflachen auftritt (boundary bias) und die Schätzung der $\hat{y}$-Werte von der Verteilung der x-Werte um x_0 abhängig ist und die Erwartungstreue der Schätzung bei nicht-symmetrischer Verteilung nicht mehr gegeben ist (Bias, Verzerrung), lassen sich diese Probleme bei der lokalen polynomialen Regression verringern.

Gemeinsam mit den bisher besprochenen Verfahren hat die lokale polynomiale Regression, dass davon ausgegangen wird, dass nicht nur der interessierende x_0-Wert und dessen y-Wert Informationen über $\hat{y}$ liefern, sondern ebenfalls die Wertepaare, die in x-Richtung als „Nachbarn“ des betreffenden Wertepaares anzusehen sind, es wird also wieder mit Kategorien oder Fenstern gearbeitet. Dabei kann die Fensterbreite fest sein (in Einheiten auf der x-Achse) und die Anzahl der im Fenster liegenden Beobachtungen variabel oder die Anzahl der Beobachtungen kann fest sein, was eine variable Fensterbreite zur Folge hat. Im Folgenden wird von letzterem Fall ausgegangen, pro Fenster wird also die gleiche Anzahl an Beobachtungen betrachtet, konstant ist hiermit die bereits erwähnte Spanne $s = \frac{m}{n}$, wobei m gleich der Anzahl der Beobachtungen pro Fenster und n gleich der Gesamtzahl der Beobachtungen ist.

Vorgehen bei der lokalen polynomialen Regression

Für jeden der interessierenden x_0-Werte wird *lokal* eine eigene Regressionskurve bestimmt, die in das jeweilige Fenster gelegt wird (deshalb *lokale* polynomiale Regression). Diese Kurve kann einer Geraden folgen, aber ebenso gut auch x-Terme mit höheren Exponenten beinhalten (deshalb lokale *polynomiale* Regression), so dass zum Beispiel pro Fenster eine Parabel eingepasst wird. Die Regressionskurve wird generell nach dem Kriterium der kleinsten gewichteten Quadrate ermittelt, so dass folgender Ausdruck minimiert wird:

$$\sum_{i=u}^{o} \omega_i^2 e_i^2 = \sum_{i=u}^{o} \omega_i^2 (y_i - \hat{y}_i)^2 = \min ;$$

wobei der Index i kontinuierlich die x-Werte anzeigt, die sich von der unteren Grenze des Fensters x_u bis zum x-Wert an der oberen Grenze x_o bewegen.

Die grau hinterlegten Teile des zu minimierenden Ausdrucks markieren den Unterschied zur einfachen Kleinstquadrateregression. Hinter ω_i verbirgt sich das Gewicht, mit dem die Abweichung des $\hat{y}$ - vom y-Wert einer Person i für die Einpassung einer Regressionskurve gewichtet wird und ist nichts anderes als das Kernel-Gewicht aus der bereits vorgestellten Kernel-Schätzung. Im Gegensatz zum Vorgehen bei dieser, wo mit dem Kernel-Gewicht *direkt* der y-Wert einer Person gewichtet wird, wird das Kernel-Gewicht in der lokalen polynomialen Regression, wie aus obigem Ausdruck hervorgeht, zur *Gewichtung des Fehlers* einer Person benutzt. Konkret bedeutet das: Je näher der x-Wert einer Person x_i am interessierenden x-Wert x_0 liegt, desto stärker wird der Fehler $(y_i - \hat{y}_i)$ bei der Einpassung der Regressionskurve in das Fenster gewichtet.

Die Abbildungen 8 a-d sollen eben erläutertes Vorgehen am Beispiel Kreativität noch einmal verdeutlichen. Der interessierende x-Wert befindet sich hierbei an der Stelle 90 und es wurde ein lokaler linearer Fit (also eine lokale Regressionsgerade) gewählt. Die Kernel-Gewichte bestimmen sich nach der trikubischen Gewichtsfunktion.

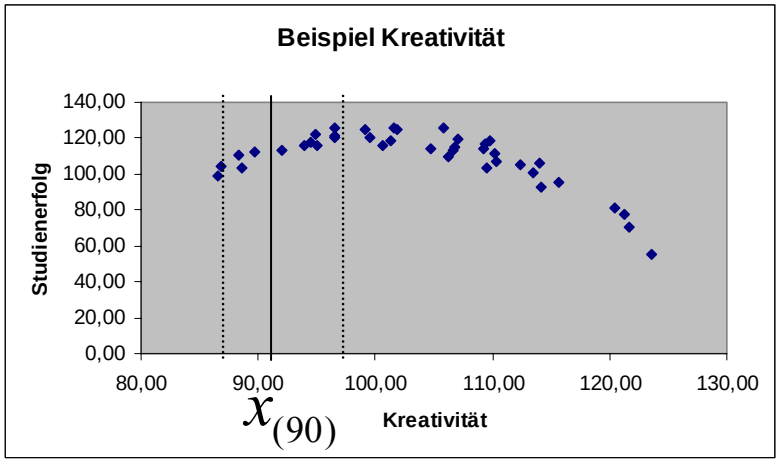


Abbildung 8a): Das Fenster schließt die m=8 nächsten Nachbarn des interessierenden Wertes $x_{(90)}$ ein.

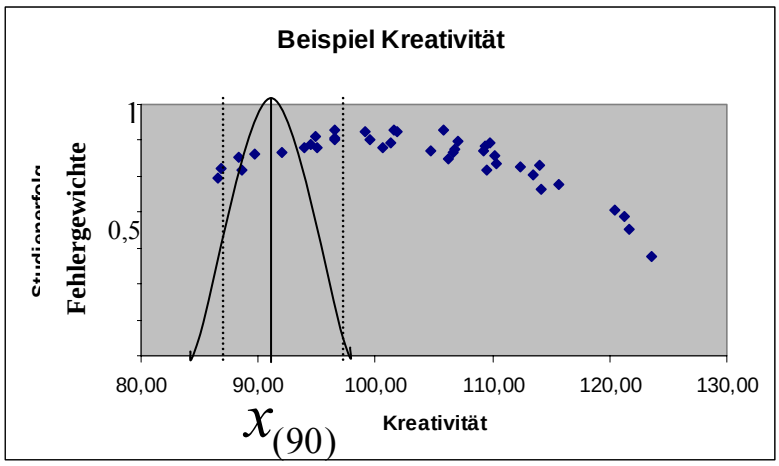


Abbildung 8b): Die trikubischen Gewichte der Fehler als Funktion der entsprechenden Kreativitätswerte.

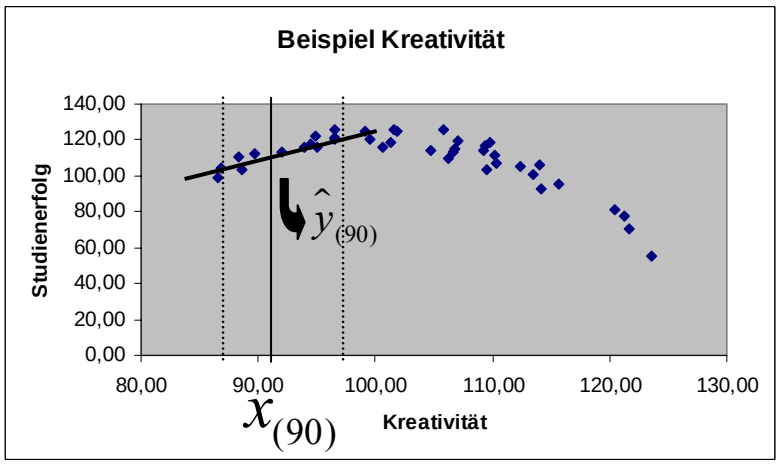


Abbildung 8c): Gezeigt wird die lokale Gewichtete-Kleinsquadrateregressionsgerade, die den geschätzten $\hat{y}_{(90)}$ -Wert produziert.

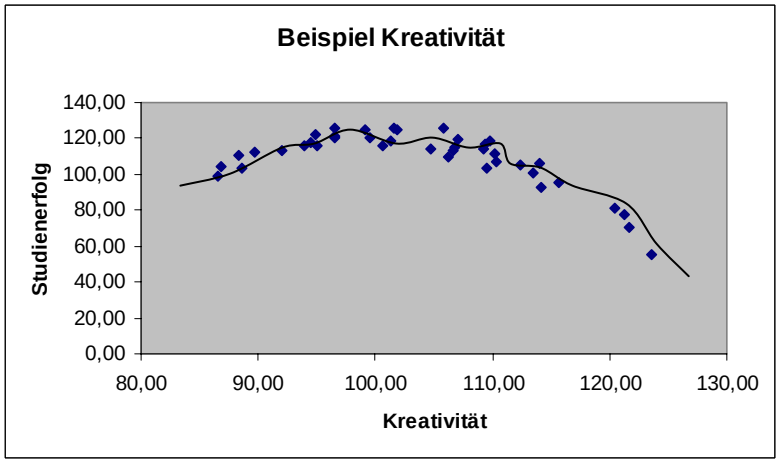


Abbildung 8d): Die $\hat{y}$ -Werte aller interessierenden x-Werte wurden zur Regressionskurve verbunden.

Wie die Kurve letztendlich aussieht ist von mehreren Faktoren abhängig. Im Prozess der lokalen polynomialen Regression ergibt sich für den Anwender an einigen Punkten die Möglichkeit, zwischen verschiedenen Alternativen auszuwählen.

Zum einen muss von Anwender bestimmt werden, welche Kernel-Gewichtsfunktion zur Gewichtung der Fehler zum Einsatz kommen soll. Auf die verschiedenen Möglichkeiten wurde bereits weiter oben eingegangen.

Des Weiteren muss entschieden werden, welchen Grades die lokale Regressionskurve sein soll. Hierzu ist zu sagen, dass höhere Exponenten von x einen flexibleren Fit produzieren (der Bias nimmt ab), gleichzeitig variieren die geschätzten $\hat{y}$-Werte jedoch auch stärker, was eine weniger glatte Kurve zur Folge hat. Generell sind ungerade geraden Exponenten vorzuziehen, da nur für ungerade Exponenten gilt, dass die Schätzung der einzelnen $\hat{y}$-Werte von der Verteilung der x-Werte um x_0 unabhängig ist, es besteht also ein Vorteil bezüglich eines geringeren Bias.

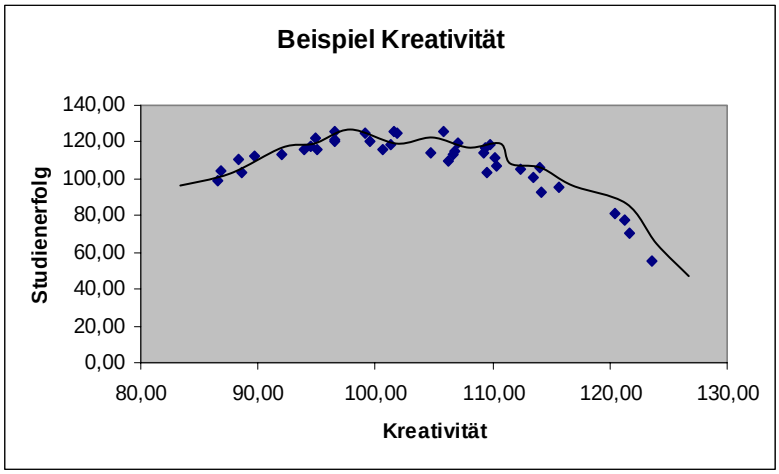
Außerdem ist die „Glattheit“ der Kurve davon abhängig, wie groß die Spanne, also die Anzahl der Beobachtungen pro Fenster ist. Generell gilt, dass größere Spannen eine glattere Kurve produzieren, jedoch besteht auch die Gefahr einer „Überglättung“. Deswegen ist es wichtig, die verwendete Spanne sorgfältig auszuwählen. Möglichkeiten hierzu werden im nächsten Abschnitt vorgestellt.

Die Auswahl der Spanne

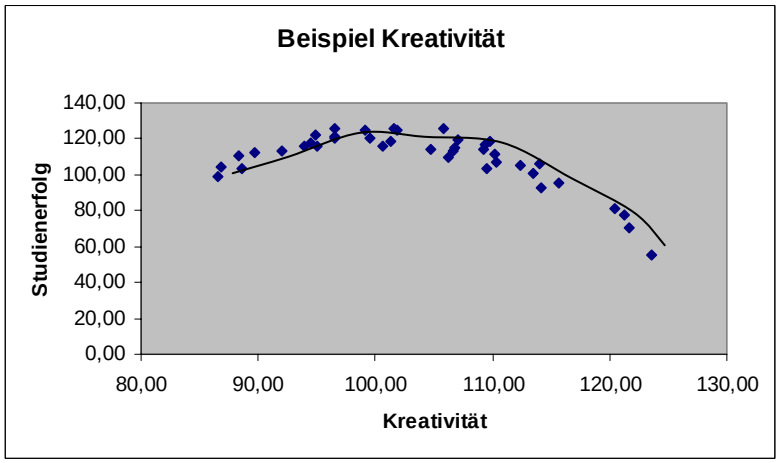
Visueller Ansatz

Im Beispiel wurde eine relativ kleine Spanne von s=8/40=0,2 verwendet, dementsprechend „holprig“ ist die Kurve. Die einfachste Möglichkeit, die Spanne auszuwählen, entspricht einem visuellen Ansatz: Die Regressionskurven, die durch verschiedene Spannen produziert worden sind, werden verglichen und man entscheidet sich für die Kurve, die bei kleinstmöglicher Spanne (da kleine Spannen einen geringeren Bias bedeuten) einen genügend glatten Eindruck macht. Das Urteil über die passende Spanne ist somit ein subjektives, da es aufgrund der möglichen Vielfalt der Kurvenform keine Richtlinien darüber geben kann, was „genügend glatt“ bedeutet. Nichtsdestotrotz ist das visuelle Verfahren ein sehr nützliches, wie aus den folgenden Abbildungen deutlich werden sollte.

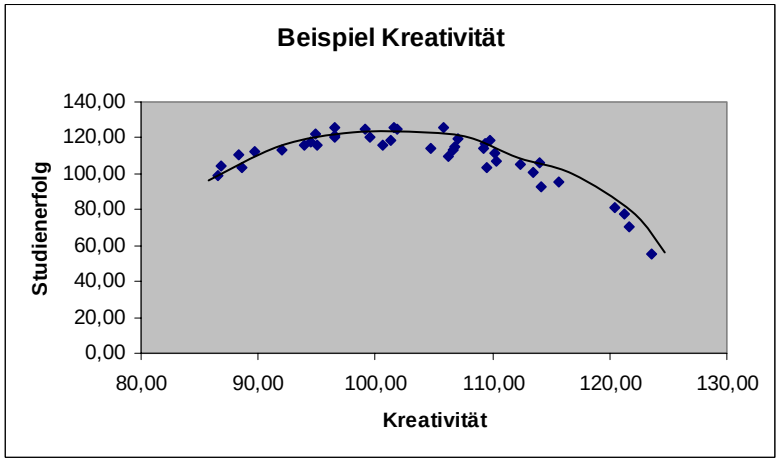
s=0,2

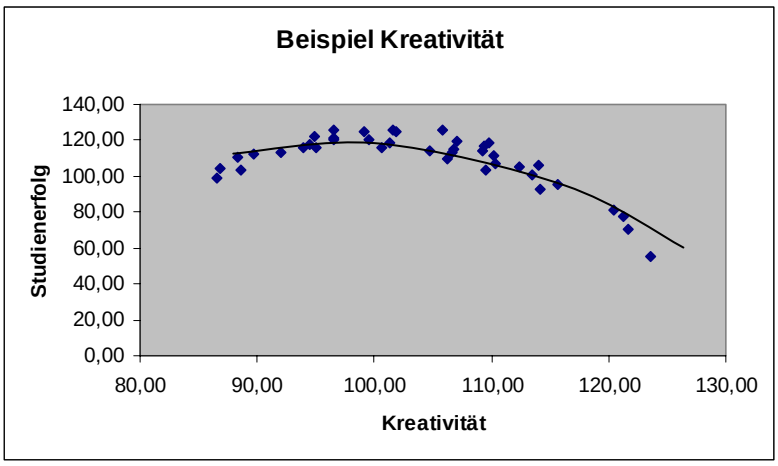

Beispiel Kreativität
Studienerfolg
140,00
120,00
100,00
80,00
60,00
40,00
20,00
0,00
80,00
90,00
100,00
110,00
120,00
130,00
Kreativität

s=0,5


Beispiel Kreativität
Studienerfolg
140,00
120,00
100,00
80,00
60,00
40,00
20,00
0,00
80,00
90,00
100,00
110,00
120,00
130,00
Kreativität

s=0,7


Beispiel Kreativität
Studienerfolg
140,00
120,00
100,00
80,00
60,00
40,00
20,00
0,00
80,00
90,00
100,00
110,00
120,00
130,00
Kreativität



s=0,9

Abbildung 9: Die verschiedenen Abbildungen zeigen die Veränderung des Aussehens der Kurve bei größer werdender Spanne. Wenn man nach einer "genügend glatten" Kurve bei möglichst kleiner Spanne sucht, würde man wahrscheinlich die Kurven mit den Spannen s=0,5 oder s=0,7 wählen.

Aufgrund dieser Abbildungen würde ein Großteil der Anwender die Spanne zwischen 0,5 und 0,7 wählen. Bei einer Spanne von 0, wo also jedes Fenster nur den interessierenden x_0-Wert beinhalten würde, wäre die Regressionskurve die Verbindung aller Punkte. Bei einer Spanne von 1 hätte die Regressionskurve die Form des gewählten lokalen Fits (also in unserem Fall eine Gerade), wobei der Fit allerdings an jeder Stelle (und damit auch insgesamt) ein globaler wäre, da ja immer alle Werte zur Schätzung eines $\hat{y}$-Wertes herangezogen würden.

Mathematischer Ansatz

Wie bereits besprochen ist die Einpassung einer Regressionskurve in der non-parametrischen immer durch Abwägen zwischen Bias und Glätte der Kurve gekennzeichnet. Zur Erinnerung: Der Bias bezieht sich auf mangelnde Erwartungstreue der einzelnen $\hat{y}$-Schätzungen und nimmt mit schrumpfender Fensterbreite (also auch mit schrumpfender Spanne) ab. Die Glätte der Kurve wächst mit sinkender Variabilität der $\hat{y}$-Werte, was durch große Kategorienbreiten bzw. große Spannen erreicht wird. Die Spanne ist also in der Lage, die beiden relevanten Größen zu beeinflussen. Mathematisch kann ein Wert für s gefunden werden, für den die Abwägung zwischen beiden für die vorliegenden Daten optimal ist.

Um dies zu verdeutlichen, wird zunächst der mean squared error of estimation MSE eines einzelnen $\hat{y}$-Wertes dargestellt, welcher sich genau aus den eben besprochenen Größen zusammensetzt, nämlich aus dem Bias und der Varianz des $\hat{y}$-Wertes:

$$MSE_{\hat{y}_i} = bias^2_{\hat{y}_i} + s^2_{\hat{y}_i}$$

Die Forderung nach einem optimalen Wert für s ist also gleichbedeutend mit der Forderung nach einem MSE, der minimal wird, denn dann sind sowohl Bias als auch Varianz so klein wie möglich (da ihre Summe ein Minimum ergibt). Da sowohl Bias als auch Varianz abhängig von der Spanne sind, lässt sich ein Wert für s finden, für den der MSE eines einzelnen $\hat{y}$-Wertes minimal wird.

Da man jedoch die optimale Spanne für alle Schätzungen der $\hat{y}$-Werte sucht, ist es notwendig, die MSEs aller $\hat{y}$-Werte zusammenzufassen. Das kann zum Beispiel über den average squared error ASE passieren, der die einzelnen MSEs zusammenfasst:

$$ASE = \frac{\sum_{i=1}^{n}\left[\hat{y}_i(s) - y_i\right]^2}{n}$$

Aus der zweiten Schreibweise wird deutlich, dass jede Spanne s für einen bestimmten Wert y_i einen anderen Wert $\hat{y}_i$ produziert und sich damit der ASE verändert. Allerdings gibt es eben einen Wert für s, an dem der ASE minimal wird, welcher die optimale Balance zwischen Bias und Varianz bewirkt. Die rechnerische Bestimmung dieses Wertes besteht im Finden des Minimums der Funktion des Durchschnittsfehlers in Abhängigkeit von der Spanne. Abbildung 10 zeigt die entsprechende Funktion.

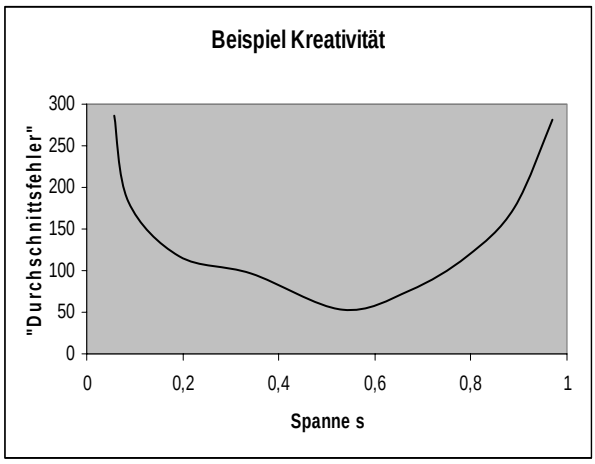


Abbildung 10: "Durchschnittsfehler" als Funktion der Spanne s für das Beispiel Kreativität.

Wie ersichtlich wird, liegt das Minimum des „Durchschnittsfehlers“, wie durch den visuellen Ansatz schon deutlich wurde, nahe einer Spanne von s=0,5, welche somit als optimal angesehen wird.

Jetzt, da das generelle Vorgehen bei der lokalen polynomialen Regression bekannt ist, soll noch kurz auf einen Sonderfall eingegangen werden, das so genannte *lowess*- Verfahren.

*Lo*cally *We*ighted *S*catterplot *S*moothing *(Lowess)*

Diese Methode stellt im Prinzip eine Erweiterung der lokalen polynomialen Regression und gleichzeitig das am besten verfügbare Verfahren zur non-parametrischen Regression dar (Fox, 2000a). Es berücksichtigt, dass die Möglichkeit von Ausreißern in den Daten besteht, dass also eventuell unübliche Werte auftreten, die die Regressionskurve entgegen der „wahren“ Beziehung beeinflussen können. Die Abbildungen 11a und b sollen kurz anhand einer linearen Kleinstquadrate-Regressionsgeraden zeigen, wie Ausreißer sich auf die Kurve auswirken können.

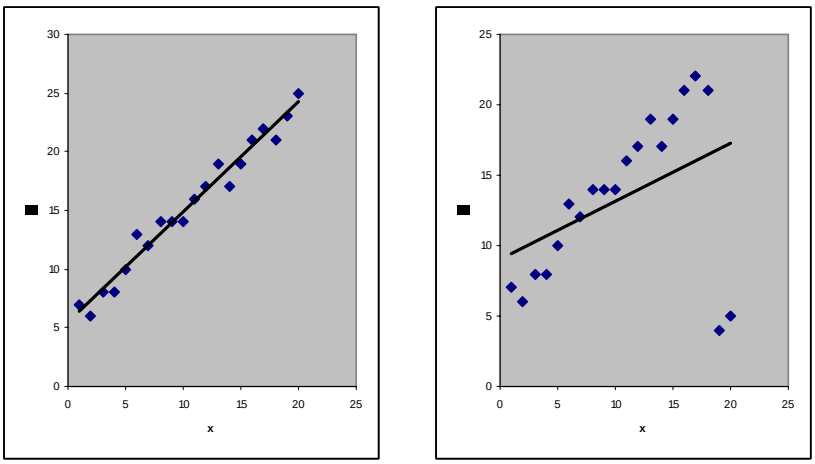


Abbildung 11: a) Die Regressionsgerade passt fast perfekt in die Form des Punkteschwarms b) Zwei Ausreißer bewirken, dass die Regressionsgerade sich dramatisch verändert: Anstatt den „Mehrheitstrend“ wiederzugeben berücksichtigt die Gerade die Lage der Ausreißer.

Durch die Hinzunahme eines weiteren Gewichts in den lokal zu minimierenden Ausdruck ist das Lowess-Verfahren weniger anfällig gegenüber Ausreißern als die einfache lokale polynomiale Regression.

Das Vorgehen beim Lowess-Verfahren ist in den Fenstern um die interessierenden x-Werte iterativ, es gliedert sich also in mehrere Durchgänge. Begonnen wird im ersten Durchgang mit einer einfachen lokalen polynomialen Regression, die in jedem Fenster die Summe der Quadrate der mit den Kernel-Gewichten gewichteten Fehler $\sum_{i=u}^{o} \omega_i^2 e_i^2$ minimiert. Aus dieser resultieren die ersten $\hat{y}$-Werte. Mit diesen ist man in der Lage, die Fehler $e_i = y_i - \hat{y}_i$ zu berechnen. Für diese Fehler e_i werden nun Gewichte W_i berechnet, die in der Regel zwischen 0 und 1 variieren, wobei kleinere Fehler stärker gewichtet werden, ist der Fehler 0, so wird das Gewicht 1. Abbildung 12 zeigt zwei gebräuchliche Gewichtungsfunktionen für die „Robustheits“-Gewichte W_i.

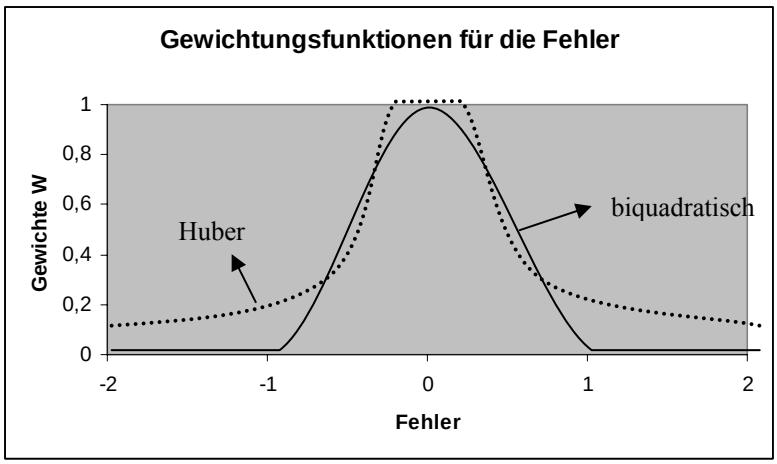


Abbildung 12: Zwei Funktionen zur Gewichtung der Fehler e_i: Die biquadratische Funktion (durchgezogene Linie) und die Huber-Funktion (gestrichelte Linie).

Beim Lowess-Verfahren, das erstmalig von Cleveland (1979, zitiert nach Fox, 2000a) vorgestellt wurde, werden neben trikubischen Kernel-Gewichten biquadratische Fehlergewichte eingesetzt.

Nach der Bestimmung der „Robustheits"-Gewichte W_i wird im zweiten Durchgang erneut eine lokale polynomiale Regression berechnet, allerdings wird dann pro Fenster folgender Ausdruck minimiert:

$$\sum_{i=u}^{o} \omega_i^2 W_i^2 e_i^2$$

Die „Robustheits"-Gewichte W_i bewirken, dass sich jede Regression innerhalb eines Fensters stärker an den Werten orientiert, die im ersten Durchgang kleine Fehler produziert haben. Da Ausreißer in der Regel große Fehler produzieren, werden diese somit weniger berücksichtigt.

Aus dem zweiten Durchgang gehen neue $\hat{y}$-Werte und neue Fehler $e_i = y_i - \hat{y}_i$ hervor, für die wiederum Gewichte berechnet werden, die dann im dritten Durchgang in den zu minimierenden Ausdruck eingehen.

Dieses Vorgehen wird solange wiederholt, bis sich die $\hat{y}$-Werte aus zwei aufeinander folgenden Durchgängen nicht mehr (oder kaum noch) unterscheiden, in der Regel reichen zwei bis vier Durchgänge aus.

Abbildung 13 zeigt, wie sich die non-parametrische Regressionskurve bei einer einfachen lokalen polynomialen Regression von einem (ins Beispiel neu eingefügten) Ausreißer beeinflussen läst und wie sich der Fit mit der Lowess-Methode nach vier Iterationen verbessert.

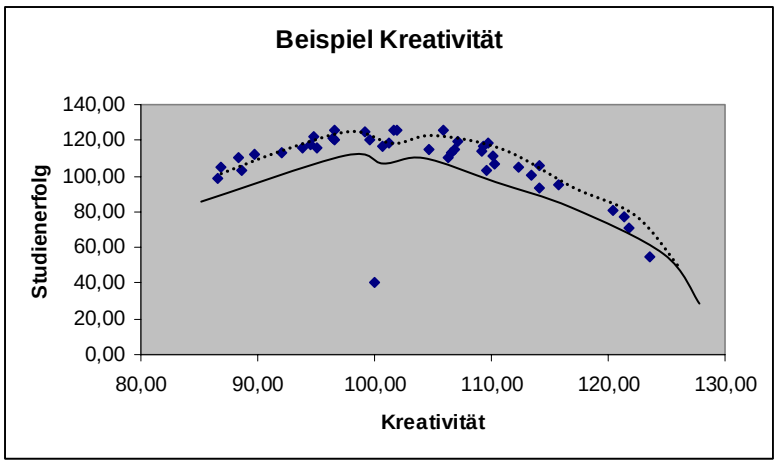


Abbildung 13: Der Fit der Regressionskurve nach einer einfachen lokalen polynomialen Regression (durchgezogene Linie) und nach Anwendung des Lowess-Verfahrens mit vier Iterationen (gestrichelte Linie). Die Spanne beträgt für beide s=0,5.

Statistische Inferenz für die lokale polynomiale Regression

In der parametrischen Regression interessieren vor allem die Regressionskoeffizienten. Die statistische Inferenz bezieht sich dann natürlich auf diese Koeffizienten, meist in Form von Konfidenzintervallen oder Hypothesentests. Im Gegensatz dazu, gibt es bei der non-parametrischen Regression keine Regressionskoeffizienten. Hier ist das zentrale Element die Schätzung der Regressionsfunktion, auf die sich auch die Inferenz direkt fokussiert.

Anwendungen der non-parametrischen Regression mit einem Prädiktor haben als Ziel oft nur das optische Smoothing eines Scatterplots. In diesen Fällen ist die Inferenzstatistik eher von geringer Bedeutung. Interessanter wird die Inferenzstatistik bei der non-parametrischen multiplen Regression. Hier wird dennoch auf die Inferenzstatistik bei der non-parametrischen Einfachregression eingegangen, vor allem auf Konfidenzhüllenbildung, Hypothesentestung und auf weitere alternative Methoden.

Konfidenzhüllen

Bei der non-parametrischen Regression wird für die Regressionsfunktion eine Konfidenzhülle gebildet. Hierzu muss man durch eine lokal gewichtete kleinste Quadrate Regression von y über die Werte von x die angepassten Werte $\hat{y}_i = \hat{y} \mid x_i$ errechnen. Der angepasste Wert kann somit auch als eine gewichtete Summe der Beobachtungen y_i dargestellt werden ($\hat{y}_i = \sum_{j=1}^{n} s_{ij} y_j$), wobei die Gewichte s_{ij} Funktionen der Werte von x darstellen. Für den Trikubischen Kernel bedeutet dies zum Beispiel, dass $s_{ij}=0$ für Beobachtungen außerhalb der Nachbarschaft des Blickpunktes x_i. Man geht davon aus, dass die Werte von y_i unabhängig verteilt sind und die gleiche Varianz σ^2 haben.

Nun muss eine Schätzung von σ^2 erfolgen, hierzu wird, wie bei der linearen Einfachregression, die Fehlervarianz s^2 errechnet, wobei die Summe der Fehler $\varepsilon_i = y_i - \hat{y}_i$ durch die Anzahl der Fehlerfreiheitsgrade geteilt wird. Die Fehlerfreheitsgrade df_{res} werden ermittelt durch $n\text{-}df_{mod}$, wobei df_{mod} die Anzahl der Freiheitsgrade des Modells darstellt.

Im Anschluss werden 95%-ige Konfidenzintervalle für jeden geschätzten Wert $\widehat{y_i}$ errechnet und verbunden. Die Verbindung der einzelnen Konfidenzintervalle für $x=x_1,x_2,\ldots,x_n$ produziert ein punktweises 95%-igen Konfidenzband, auch Konfidenzhülle genannt, für die Regressionsfunktion. In einer solchen punktweisen Konfidenzhülle bezieht sich die Konfidenzaussage nur auf die einzelnen x_i. Eine simultane Hülle zu konstruieren für die Regressionsfunktion als Ganzes ist ein schwieriges und unpraktisches Unterfangen.

Das Vorgehen bei der Konstruktion einer Konfidenzhülle ist zwar einfach, aber es ist nicht ganz korrekt, aufgrund des Bias in $\hat{y}$ als Schätzer des Erwartungswertes von y. Wenn die Wahl der Spanne und des Grades des lokal-polynomialen Schätzers allerdings vernünftig ist, dann dürfte der Bias klein sein. Der Bias hat die folgenden Konsequenzen:

- s^2 wird systematisch überschätzt, dies führt zu einer Überbewertung der Fehlervarianz und der Konfidenzintervall wird zu breit.
- Das Konfidenzintervall wird im Durchschnitt an der falschen Stelle zentriert.

Diese Fehler gleichen sich oft von selbst aus, aber weil $\hat{y}$ doch noch biased ist, ist es genauer, wenn man die Hülle um die Regressionskurve als Variabilitätsband bezeichnet, und nicht als Konfidenzband.

Hypothesentests

Analog zur Regression der kleinsten Quadrate, besteht die Hypothesentestung in der non-parametrischen Regression darin, den F-Test durchzuführen. Der F-Test prüft eine angenommene Null-Hypothese, also den nicht bestehenden Zusammenhang zwischen Variablen. Er wird mit folgender Formel errechnet:

$$F = \frac{(TSS - RSS)/df_{\text{mod}} - 1)}{RSS/df_{res}},$$

wobei $df_{res}=n\text{-}df_{mod}$. RSS bedeutet Fehlerquadratsumme.

Die Non-linearität kann geprüft werden, indem das non-parametrische Regressionsmodell gegen das lineare Einfachregressionsmodell kontrastiert wird. Die beiden Modelle gehen auseinander hervor, weil eine lineare Beziehung ein besonderer Fall der allgemeinen, potenziell nichtlinearen, Beziehung ist. Wir nennen die Fehlerquadratsumme des linearen Modells RSS_0 und die des allgemeinen non-parametrischen Modells RSS_1. So erhalten wir die Formel

$$F = \frac{(RSS_0 - RSS_1) / df_{\text{mod}} - 2)}{RSS_1 / df_{res}}$$

Dieser Test ist konstruiert unter Berücksichtigung der Regel, dass das allgemeinere Modell, hier das non-parametrische Regressionsmodell, angewandt wird zur Schätzung der Fehlervarianz $S^2 = RSS_1/df_{res}$.

Alternative inferenzstatistische Methoden

Bootstrap Konfidenzhüllen

Das Bootstrapping ermöglicht einen Zugang zur statistischen Inferenz und basiert auf der zufälligen Wiederholungsprobennahme der Daten. Diese ist eine sehr attraktive Methode, weil sie keine starken Annahmen über die Verteilung benötigt und weil sie in Fällen angewandt werden kann, in denen analytische Ergebnisse schwer zu erreichen sind. Der Nachteil des Bootstrappings liegt darin, dass es sehr rechenintensiv ist und eventuell eine maßgeschneiderte Programmierung voraussetzt.

Um Bootstrapping einzusetzen, betrachten wir die Stichprobe als ganze Population, aus der n Beobachtungen aus den Daten zufällig ausgewählt werden und zurückgesetzt werden. So erhalten wir Schätzungen für die resultierende Bootstrap-Stichprobe. Dieser Vorgang wird sehr oft wiederholt und jedesmal die Schätzer für einen neuen Bootstrap kalkuliert. Es ist hier sehr wichtig, dass die Beobachtungen zurückgesetzt werden, sonst würde eine reine Kopie der ursprünglichen Stichprobe entstehen. Die Konsequenz des Zurücksetzens der Daten ist, dass einige Beobachtungen typischerweise mehr als einmal in der Bootstrap-Stichprobe auftauchen, andere jedoch gar nicht.

Die Analogie des Bootstrappings ist, dass die Bootstrap-Stichproben vergleichbar sind mit der ursprünglichen Stichprobe in dem Maße, indem die ursprüngliche Stichprobe vergleichbar ist mit der Population. Insofern wird erhofft aus den Bootstrap-Schätzer etwas über die Verteilung der Schätzer zu lernen, die aus der ursprünglichen Stichprobe resultierten.

Die Methode, die hier beschrieben wurde, geht von einer unabhängigen zufälligen Stichprobe aus der Population aus. Wenn ein anderes Stichprobenmodell gewählt wurde, sollte dies in der Wahl der Bootstrap-Stichproben gespiegelt werden. Die wichtigste Bedingung der Wahl der Bootstrap-Stichproben ist, dass sie in Bezug auf die Wahl der ursprünglichen Stichprobe aus der Population parallelisiert werden, eine logische Folge der Bootstrap-Analogie.

Für Regressionsdaten entnimmt man beim Bootstrapping (mit Zurücksetzen) immer x,y Paare, $\{x_1^*,y_1^*\},\{x_2^*,y_2^*\},\ldots,\{x_n^*,y_n^*\}$, wobei die Sternchen daran erinnern sollen, dass z.B.

die Beobachtung $\{x_1^*, y_1^*\}$ nicht unbedingt die erste Beobachtung in der ursprünglichen Stichprobe war, sondern die erste zufällige Probennahme ist.

Dieses Vorgehen wird B mal wiederholt, wobei jedes mal eine Bootstrap-Stichprobe gezogen wird und die non-parametrische Regression rekalkuliert wird. Die B-te Stichprobe enthält die angepassten Werte $\hat{y}_{b1}^*$, $\hat{y}_{b2}^*$,..., $\hat{y}_{bn}^*$. Um Konfidenzintervalle bilden zu können, muss B sehr groß sein, mindestens 1000.

Eine Möglichkeit das Konfidenzintervall für E(y) an der Stelle der Regressionskurve x_i zu bilden, ist zunächst die B Bootstrap Replikate von $\hat{y}_i^*$ zu ordnen von kleinstem Wert bis größtem Wert: $\hat{y}_{(1)i}^*$, $\hat{y}_{(2)i}^*$,..., $\hat{y}_{(B)i}^*$. Wenn wir ein 95%-iges Konfidenzintervall erhalten wollen, dann liegen die Endpunkte des perzentilen Konfidenzintervalls für E(y) bei $\hat{y}_{(25)i}^*$ und $\hat{y}_{(975)i}^*$, diese markieren die 2,5 und 97,5 Perzentile der Verteilung $\hat{y}_i^*$. Diese Rechnung erfolgt an jeder Stelle, an der die Regression evaluiert ist, um eine punktweise 95%-ige Konfidenzhülle für die non-parametrische Regression zu konstruieren.

Randomisierung

Randomisierungstests ähneln dem Bootstrap. Sie sind in der non-parametrischen Regression bei bestimmten Hypothesen anwendbar, wie bei der Nullhypothese. Die Nullhypothese geht von einer nicht bestehenden Beziehung aus zwischen y und x. Wenn die Nullhypothese bestätigt wird, bedeutet dies, dass der Erwartungswert von y an jedem x gleich sein muss.

Nun errechnen wir eine Teststatistik, wie den T-Wert für die Nullhypothese, der den Grad der Abweichung der Daten von H_0 schätzt. Wir nennen den beobachteten Wert dieser Statistik T^*. Die Daten werden jetzt vertauscht, indem beliebige Paare von den vorhandenen x- und y-Werten gebildet werden, um einen neuen Datensatz $\{x_i, y_i^*\}$ zu bilden. Die non-parametrische Regression wird an die vertauschten Daten angepasst, und der Wert der Teststatistik neu errechnet.

Wenn n sehr klein ist, können alle M=n! Permutationen der Daten konstruiert werden, für die jedesmal die Regression und die Teststatistik neu kalkuliert wird. Die Verteilung von T über die vertauschten Datensätze entspricht der empirischen Stichproben-Verteilung der Teststatistik, angenommen die Nullhypothese gilt. Der p-Wert des Permutationstests ist deshalb der Anteil der Teststatistiken in der empirischen Stichproben-Verteilung von T, die größer sind als der beobachtete Wert T^*.

In den Fällen, in denen n nicht sehr klein ist, ist es unpraktisch alle möglichen vertauschten Paare von x- und y-Werten zu erheben. Eine wirksame Alternative besteht darin, eine Stichprobe zu nehmen, die aus einer relativ großen Anzahl von Permutationen besteht, um dann den Schätzwert des p-Wertes für den Permutationstest zu kalkulieren. Dieses Vorgehen nennt man Randomisierungstest.

Regression Splines

Regression Splines stellen neben der lokalen polynomialen Regression einen weiteren möglichen Ansatz zur non-parametrischen Regression dar. Wieder geht es darum, in eine Punktewolke eine möglichst gut passende Regressionskurve hineinzulegen, die der entsprechende Regressions-Spline darstellen soll, dafür wird wiederum das Streudiagramm in einzelne Abschnitte unterteilt. Regressions-Splines sind stückweise polynomiale Funktionen, die dahingehend beschränkt werden, dass sie sich an den Kategoriengrenzen, die hier knots genannt werden, glatt einander annähern (siehe Abbildung 14 a und b).

In jeden Abschnitt wird nun eine eigene polynomiale Kurve hineingelegt, von der angenommen wird, dass sie die Daten in diesem Abschnitt gut repräsentiert. Dies impliziert, dass man für jeden Abschnitt eine solche, explizit formulierte Vorstellung hat, das Verfahren produziert also am Ende einen vollständig parametrischen Fit.

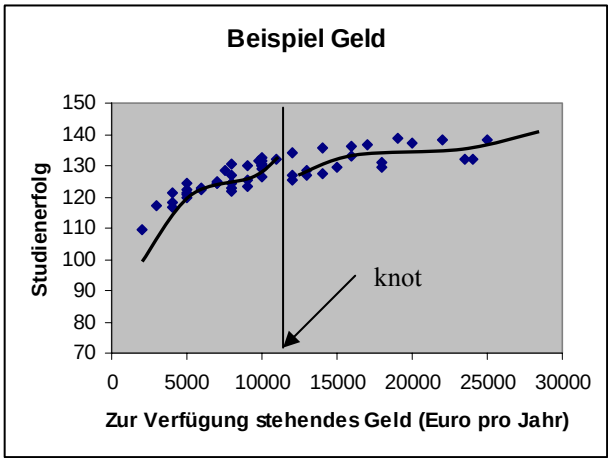


Abbildung 14a): Zunächst wird in jedes Fenster eine eigene polynomiale Funktion eingepasst (hier jeweils eine kubische).

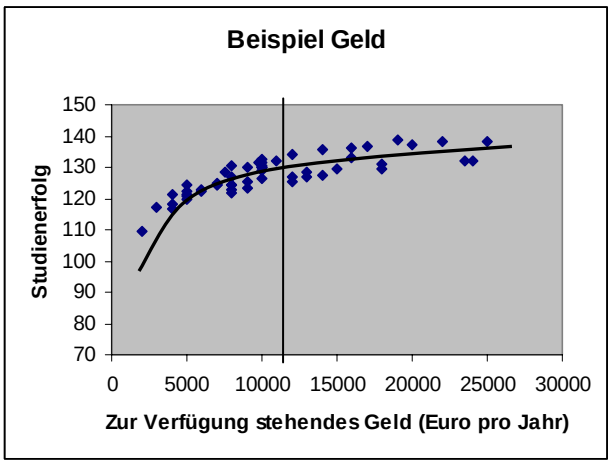


Abbildung 14b): Die Kurven werden dazu gebracht, sich in den knots anzunähern.

Abbildung 14 verdeutlicht noch einmal das Vorgehen bei der Entwicklung eines Regressions-Splines am Beispiel zur Beziehung zwischen zur Verfügung stehendem Geld und Studienerfolg.

Bei der Verwendung von Regressions-Splines nähern sich die Kurven in den knots glatt an. In Abbildung 14 b) wird ein natürlicher kubischer Spline dargestellt, der den beiden Funktionen nicht nur auferlegt, sich in den knots glatt anzunähern, sondern zusätzlich knots an den Datengrenzen und einen linearen Fit darüber hinaus annimmt, um die Kurve auch an den Extrema der Daten zu glätten.

Die Entwicklung eines Regression-Splines benutzt also die Methode eines linearen Modells und ist somit rechnerisch nur durch die Beschränkung der einzelnen polynomialen Funktionen, sich in den knots einander anzunähern etwas komplexer. Die praktische Schwierigkeit für den Anwender ergibt sich auch hier wieder über Kategorienbildung und ist die Frage, wie viele knots gewählt und wo genau sie platziert werden sollen. Auf dieses Problem soll hier jedoch nicht mehr eingegangen werden.

Des weiteren gibt es auch die Möglichkeit im Rahmen der Spline-Berechnung zwischen zwei knots jeweils einen non-parametrischen Fit zu erzeugen, wobei diese dann in den knots ebenfalls zum Konvergieren gebracht werden.

Non-parametrische Regression und Datenanalyse

Der Scatterplot ist einer der wichtigsten datenanalytischen statistischen Graphen, der in sehr unterschiedlichen Bereichen der Datenanalyse anwendbar ist. Hier wird eingegangen auf das Nutzen des Scatterplots in Bezug auf die non-parametrische Regression, nämlich inwieweit der Scatterplot hilfreich ist beim Aufdecken von Non-Linearität. Die Möglichkeit, dass Non-Linearität herrscht, ist allgegenwärtig in der Regressionsanalyse. Die non-parametrische Regression bietet eine Möglichkeit mit der Non-Linearität umzugehen. Eine Alternative besteht darin, eine lineare Regression für die Daten einzusetzen, die Daten also zu transformieren.

Potenztransformationen

Die Transformation von Daten erfolgt durch das Ersetzen der Variablen x mit der Potenz x^p. Wenn p=2, dann wird die Variable ersetzt mit ihrer Quadratzahl x^2; wenn p=-1, wird die Variable mit ihrer Umkehrfunktion $x^{-1}=1/x$ ersetzt; wenn p=1/2, wird die Variable mit ihrer Wurzel $x^{\frac{1}{2}} = \sqrt{x}$ ersetzt usw. Die einzige Ausnahme ist, dass p=0 die logarithmische Transformation log*x* beschreibt.

Transformationen dieser Art sind nur durchführbar, wenn die Werte von x positiv sind. Um diesem Problem aus dem Weg zu gehen, wird zunächst eine Konstante c zu allen Werten vor der Transformation addiert. Ein anderes Problem ist das Benutzen der Umkehrfunktion,

denn hier verändert sich die Reihenfolge der x-Werte nach der Transformation. Um die ursprüngliche Reihenfolge beizubehalten, wird $x \rightarrow -x^p$, wenn p negativ ist.

Potenztransformationen von x oder y können helfen, einen linearen Zusammenhang aus einem non-linearen Zusammenhang, der sowohl einfach als auch monoton ist, zu machen.

- Ein Zusammenhang ist einfach, wenn die Kurve weich verläuft, ohne die Richtung zu ändern.
- Ein Zusammenhang ist monoton, wenn y strikt zu- oder abnimmt mit x.

Obwohl non-lineare Zusammenhänge, die entweder nicht einfach oder nicht monoton sind, nicht durch Potenztransformation linearisiert werden können, können hier andere Formen der parametrischen Regression angewandt werden.

Fazit

Abschließend bleibt noch zu sagen, dass die Vorteile der non-parametrischen Regression groß sind. Mit ihr können Zusammenhänge erkannt werden, ob lineare oder non-lineare, ohne dabei Gefahr zu laufen, einen bestimmten Zusammenhang zu vermuten und damit die Resultate in gewisser Weise zu verfälschen bei falscher Vermutung des Zusammenhangs. Diese Methode ist vielfach anwendbar, da Scatterplots bei der Datenanalyse allgegenwärtig sind, und weil Scatterplot-Smoothing Programme mehrfach verfügbar sind. Außerdem ist es einfach mit der Potenztransformation non-lineare Zusammenhänge in lineare umzuwandeln, um so mehrere Zusammenhänge miteinander zu vergleichen. Die non-parametrische Einfachregression wird dennoch am besten als zusätzliche Methode zu anderen statistischen Methoden angesehen, da sie auch Fehler bergen kann, z.B. den Bias. Sie ist der Vorläufer der non-parametrischen multiplen Regression, die den Zusammenhang zwischen einer abhängigen Variablen und mehreren Prädiktorvariablen entdecken kann.

Literatur

Fox, J. (2000a). *Nonparametric Simple Regression: Smoothing Scatterplots.* Sage Universitiy Papers Series on Quantitative Applications in the Social Sciences, 07-130. Thousand Oaks, CA: Sage.

Einführung in die Survival Analysis

Jasmin Honold

Überblick

Nachfolgend soll eine Einführung in verschiedene Methoden zur statistischen Analyse von Daten gegeben werden, welche die Zeitdauer bis zum Eintreten eines bestimmten Ereignisses beschreiben, z.B. die Lebensdauer von technischen Systemen oder Lebewesen. Daher wird das Gebiet, das ursprünglich innerhalb der Biostatistik entwickelt wurde, auch als "Survival Analysis" bezeichnet. Dieses Verfahren erlaubt es, die Lebensdauerverteilung auch bei zensierten Daten zu schätzen. Die hierbei verwendeten Methoden sind sehr vielfältig. Sie reichen von einparametrischen Modellen über den nicht-parametrischen Kaplan-Meier-Schätzer bis zu Modellen, die sich auf die Theorie der Punktprozesse und der Martingales stützen (für einen exzellenten Überblick der verschiedenen Verfahren siehe Harrell, 2001).

In letzter Zeit sind neue Anwendungsfelder der Survival Analysis hinzugekommen und damit auch Erweiterungen der Modelle entwickelt worden. Es soll versucht werden, die Nützlichkeit der Methoden der Survival Analysis im Kontext der Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs zu beleuchten. Zudem werden einige Hinweise zur Planung und Analyse von Untersuchungen mit zensierten Datensätzen dargestellt. Daran anschließend sollen die Grundzüge der Survival Analysis ohne Annahme explanatorischer Variablen thematisiert werden.

Einleitung

Viele Untersuchungen der empirischen Wissenschaften haben zum Ziel, an einer oder mehrerer Versuchsgruppen festzustellen, ob ein bestimmtes Ereignis als Folge einer Manipulation eintritt, und wenn ja, wann ein solches eintritt. Unter der Annahme einer Normalverteilung des interessierenden Konstrukts in der Zielpopulation (in diesem Falle dem zeitlichen Verlauf zwischen einer Manipulation und einem interessierenden Ereignis) sowie der Bedingung, dass zum Zeitpunkt des Abschlusses der Datenerhebung für jeden Probanden eine definitive Aussage darüber möglich ist, wann ein solches Ereignis eingetreten ist, dienen gewöhnliche Regressionsmodelle dazu, Vorhersagen über die Zeit bis zum Eintreten des Ereignisses zu machen; eine erwartete Zeit wird dabei als die Funktion eines gewichteten Prädiktors bzw. als eine Linearkombination mehrerer Prädiktoren dargestellt.

Die Datenanalyse einer Stichprobe anhand solcher Modelle ist insbesondere dann problematisch, wenn eine Datenerhebung abgeschlossen werden muss, bevor das Ereignis

bei allen Probanden eingetreten ist. Als ein Beispiel ist hier an medizinische Untersuchungen im Kontext unheilbarer Krankheiten mit Todesfolge zu denken, in denen erhoben werden soll, inwieweit sich eine bestimmte Behandlung wie etwa eine Operation oder Medikationen auf eine verlängerte Lebenserwartung auswirken; das Zielereignis der Studie wäre hierbei der Tod, die interessierende abhängige Variable demnach die Zeit von der Behandlung bis zum Tod. Nun ist es nicht unwahrscheinlich, dass einige Probanden länger leben, als eine Studie nicht zuletzt aus ökonomischen oder zeitlich drängenden Gründen durchgeführt werden kann. Man spricht in diesem Falle von *(rechts-) zensierten* Daten, die in *Abbildung 1* veranschaulicht sind.

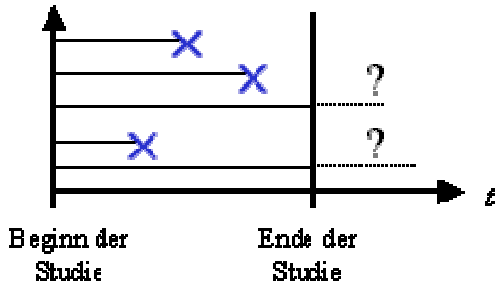


Abbildung 1: Daten von Probanden, bei denen ein interessierendes Ereignis zum Zeitpunkt des Abschlusses einer Datenerhebung nicht aufgetreten ist, werden als rechts-zensiert bezeichnet.

Nach Singer & Willett (1991) sind einige Forscher diese Problematik angegangen, indem sie ihre Daten auf unzensierte Beobachtungen gekürzt haben, betreffende Probanden als eine eigene Gruppe kategorisiert und gesondert interpretiert haben, oder aber post hoc ihre Hypothesen dahingehend geändert haben, nicht mehr den Zeitpunkt eines Ereignisses zu untersuchen, sondern lediglich festzustellen, ob ein bestimmtes Ereignis innerhalb eines definierten Zeitrahmens auftritt oder nicht. Da diese Strategien jedoch augenscheinlich mit großen Einschränkungen behaftet sind, wurden im Bereich der Biostatistik neue Methoden entwickelt, um eben solche zensierten Fälle in die Schätzungen der Zeit zwischen medizinischen Behandlungen und dem Tod miteinbeziehen zu können; Pionierarbeit wurde hierbei insbesondere von Cox geleistet (1972, zitiert nach Singer & Willett). Diese Modelle werden unter dem Begriff *Survival Analysis* subsumiert. Die Möglichkeiten der gelegentlich auch als *event history analysis* oder *hazards modeling* bezeichneten Verfahren erkannten jedoch auch Forscher anderer wissenschaftlicher Disziplinen, so dass Survival Analysis innerhalb weniger Jahre unter anderem von Ingenieuren, Ökonomen und Soziologen angewandt und weiterentwickelt wurde. So wird Survival Analysis heute nicht nur bei Fragestellungen über die Lebensdauer von technischen Geräten oder von Menschen angewandt, sondern kommt auch in anderen klinischen und epidemiologischen Studien zum Einsatz (um beispielsweise die Effektivität von Alternativinterventionen bzw. Placebomedikation zu untersuchen, Risikofaktoren bestimmter Erkrankungen zu

identifizieren oder prognostische Faktoren in Bezug auf den Verlauf von Krankheiten zu finden). Zudem findet sich dieses Verfahren auch in der psychologischen Forschung bei einer Vielzahl von Fragestellungen; beispielsweise wurde Survival Analysis angewandt, um Verhalten in Organisationen oder soziale Interaktion zu untersuchen (z.B. Allison & Liker, 1982 bzw. Fichmann, 1988, zitiert nach Singer & Willett).

Auch im Kontext der Auswahl von Studienbewerbern durch Universitäten erscheint Survival Analysis als ein geeignetes Verfahren, um die prognostische Validität verschiedener, als Prädiktoren für den Studienerfolg herangezogener Variablen zu bestimmen. Als ein Beispiel sei hier an eine retrospektive Studie zu denken, in welcher der Zusammenhang zwischen der Abiturdurchschnittsnote (und/oder einer oder mehrerer Einzelnoten) und einem oder mehrerer Kriterien für den Studienerfolg, wie etwa dem Studienabschluss und der Abschlussnoten sowie der Studiendauer, in einem Fachbereich untersucht werden soll. Um die interne Validität einer solchen Studie zu sichern, dürften nur Studierende miteinbezogen werden, die unter vergleichbaren Bedingungen studieren (wie z.B. unter der gleichen Studien- und Prüfungsordnung eines bestimmten Studiengangs an einer bestimmten Universität; im Kontext der Einführung von Bachelor- und Masterstudiengängen wäre hier zu beachten, welcher der angestrebte Studienabschluss der Studierenden ist). Da es wahrscheinlich ist, dass sich solche Bedingungen häufiger ändern, als dass alle Studienanfänger eines bestimmten Semesters ihr Studium definitiv abgeschlossen oder aber abgebrochen haben, ergäbe sich das Problem, wie mit Langzeitstudierenden umgegangen werden sollte: Würde man sich dazu entscheiden, zu einem bestimmten Zeitpunkt nur solche Studierende in die Untersuchung aufzunehmen, die ihr Studium etwa nach 14 Semestern abgeschlossen haben, gingen wichtige Informationen verloren, da Langzeitstudierende unberücksichtigt blieben. Da, wie weiter oben erwähnt, Survival Analysis auch zensierte Daten in die Berechnung der interessierenden abhängigen Variablen, in diesem Falle das gewählte Kriterium für ein erfolgreiches Studium, einbeziehen kann, erscheint sie hier als die Methode der Wahl.

Die nach dem 7. HRGÄndG (Gesetz zur Änderung des Hochschulrahmengesetzes) beschlossene Umstrukturierung zum Wintersemester 2005/2006 bei der Auswahl von Studierenden sieht vor, dass nicht mehr alle Studienanfänger der zentralen Numerus Clausus-Fächer durch die Zentralstelle für die Vergabe von Studienplätzen (ZVS) ausgewählt werden sollen, sondern 60% dieser Studienplätze durch die Universitäten selbst vergeben werden. Dadurch ist es nun ein Anliegen der Universitäten, ein optimiertes Auswahlverfahren zu entwickeln und so solche Studienanfänger auszuwählen, die mit großer Wahrscheinlichkeit zu einem erfolgreichen Studienabschluss kommen werden. Es ist zu erwarten, dass durch die Hinzunahme weiterer, mit Studienerfolg in Zusammenhang stehender und nach o.g. Gesetz zugelassener Merkmale Selektionsentscheidungen verbessert werden können. So sollte etwa versucht werden, fachspezifische Studierfähigkeitstests oder strukturierte Aufnahmegespräche zu entwickeln, die zusätzlich

zu bei der Bewerbung bereits verfügbaren Daten wie Abiturnoten oder Berufsqualifikation herangezogen werden können. Um einzuschätzen, welchen zusätzlichen Beitrag solche Variablen zur Vorhersage des Studienerfolgs leisten, muss es prospektive Studien geben. Als ein weiteres Beispiel für die Anwendung der Survival Analysis sei hier an die Evaluation eines Studierfähigkeitstests für das Studium der Psychologie zu denken, an dem alle Studienanfänger eines bestimmten Semesters anonymisiert teilnehmen sollten. Geht man nun davon aus, dass eine erfolgreich abgeschlossene Diplom-Vorprüfung als ein Kriterium für die Vorhersage des Studienerfolgs dient, könnte bereits nach zwei Jahren erhoben werden, wie viele Studenten das Grundstudium nach 4 Semestern abgeschlossen haben, und mit welcher Durchschnittsnote. Damit wäre ein Rückschluss auf die inkrementelle prognostische Validität eines solchen Tests (im Hinblick auf die Vorhersage des Studienerfolgs eines Bewerbers) möglich, die sodann gegen den Aufwand und die Kosten, die mit der Durchführung eines solchen Aufnahmeverfahrens verbunden sind, abzuwägen wäre. In diesem Falle wäre es erstrebenswert, möglichst schnelle Resultate zu erhalten, um den Test baldmöglichst überarbeiten zu können. Da viele Studenten ihr Grundstudium nicht nach 4 Semestern abschließen, lägen zensierte Daten vor, die unter Anwendung der Survival Analysis dennoch berücksichtigt würden.

Wie hier deutlich wird, ist Survival Analysis ein Verfahren, das bei der Auswahl von Studierenden durch die Hochschulen von großem Nutzen sein kann. Daher soll in der vorliegenden Arbeit nun detaillierter darauf eingegangen werden, unter welchen Rahmenbedingungen diese Methode angewandt werden kann, und was es bei der Planung von Studien, die mittels Survival Analysis analysiert werden sollen, zu beachten gilt.

Möglichkeiten und Grenzen der Survival Analysis

Als das Ziel einer Survival Analysis definieren Bull & Spiegelhalter (1997) die Schätzung der Wahrscheinlichkeit, zu verschiedenen Zeitpunkten zu überleben oder von einem fraglichen Ereignis unbetroffen zu sein, und diese Beziehung kann durch die *Überlebensfunktion* beschrieben werden. Als die abhängige Variable, respektive Reaktionsvariable, solcher Untersuchungen kann also das Zeitintervall vom Beginn der Beobachtung eines Probanden bis zum Eintreten eines kritischen Ereignisses bezeichnet werden, die oftmals *Überlebenszeit* (*survival time*) oder *Ereigniszeit* (*event time*) genannt wird. Die graphische Darstellung der Überlebensfunktion, auf die im zweiten Teil dieser Arbeit detaillierter eingegangen wird, ermöglicht eine Übersicht über erhobene Informationen, die immer in Zusammenhang mit der Zeit zu interpretieren sind.

Wie bereits eingangs erwähnt, können mit Survival Analysis zensierte Datensätze analysiert werden, wobei Bull & Spiegelhalter *Zensur* als den Verlust von Follow-Up-Untersuchungen an Probanden aufgrund anderer Ereignisse als das interessierende bezeichnen (wie hier wird im Folgenden dann von *zensierten Daten* gesprochen, wenn eine Zensierung „von rechts“ gemeint ist, d.h. auf einer Zeitachse, welche die Beobachtungsperiode einer Studie

beschreibt, kommt es zu Informationsverlust in Richtung nach rechts, also gegen Ende der Beobachtungsperiode).

Davon abzugrenzen ist *Intervall-Zensierung,* die beispielsweise vorliegt, wenn eine bestimmte Reaktion nur innerhalb eines bestimmten Intervalls gemessen werden kann. Reaktionen außerhalb dieses Bereichs sind am Ende der Skala des Messinstruments zensiert. Eine Intervall-Zensierung findet sich auch, wenn das Vorhandensein des Zielereignisses durch periodische Messungen erfasst wird (es wird je nach Fragestellung vielleicht am Ende eines jeden Semesters erfasst, ob Personen noch studieren). Wenn die Begebenheit (z.B. ein Studienabbruch) zum Zeitpunkt einer Messung eingetreten sein sollte, weiß man lediglich, dass sich die Zeit, zu der das Ereignis passiert ist, zwischen der vorangegangenen und der jetzigen Untersuchung befindet, eine punktgenaue Aussage ist jedoch nicht möglich.

Survival Analysis kann in experimentellen Studien oder in Feldstudien angewandt werden, die retrospektiv oder prospektiv sein können und in denen die Zeit kontinuierlich erfasst werden kann (beispielsweise anhand einer Kamera, die eine Verhaltenssequenz aufzeichnet) oder diskret (also mehrere Messungen zu verschiedenen Zeitpunkten; in retrospektiven Untersuchungen genügt eine einmalige Messung). Das Ziel solcher Studien kann ebenso vielfältig sein; nach Harrell (2001) ist das Verfahren der Survival Analysis – wie auch gewöhnliche Regressionsmodelle – neben einfachen Vergleichen unter anderem geeignet, um

- Wechselwirkungen verschiedener Versuchsgruppen und Interventionen zu untersuchen und zu beschreiben, anstatt Subgruppen-Analysen durchzuführen;
- die Stärke und die Art prognostischer Faktoren zu verstehen;
- (statistische) Modelle für die Effekte verschiedener Interventionen zu erstellen;
- den zeitlichen Verlauf verschiedener Interventionseffekte zu verstehen;
- vor allem in Feldstudien Stärke für die Testung von Effekten (Effektstärke) zu gewinnen und um
- ungleiche Verteilungen von Probanden zu Untersuchungsgruppen in Feldstudien auszugleichen.

Zu beachten sei hierbei, dass auch bei vollständigen, also nicht-zensierten, Datensätzen die Anwendung einer Survival Analysis gegenüber gewöhnlicher Regressionsmodelle von Vorteil sein kann, etwa wenn nicht von einer Normalverteilung eines Konstrukts in der Population ausgegangen werden kann; zudem kann eine Funktion, die in der Survival Analysis angewandt wird (*Hazard-Funktion,* die im zweiten Teil dieser Arbeit erläutert wird) zum besseren Verständnis von Überlebensmechanismen beitragen.

Neben den allgemeinen Voraussetzungen, die jeglicher statistischer Datenanalyse zugrunde liegen, sind Datensätze, die anhand von Survival Analysis untersucht werden sollen, an zwei Bedingungen gebunden:

1. Jeder Proband kann zu jedem Zeitpunkt, an dem eine Messung erfolgen soll, einem und nur einem bestimmten Status zugeordnet werden; das heißt es gibt kategoriale Merkmalsausprägungen der Probanden, die sich gegenseitig ausschließen und nicht-überlappend sind (wie beispielsweise Leben und Tod oder abgeschlossenes und andauerndes Studium).
2. Für mindestens einige der Probanden ist bekannt, wann der Übergang von einem in den anderen Status stattfindet (also wann Probanden sterben bzw. ihr Studium erfolgreich beenden).

Während in experimentellen Studien mit randomisierten Versuchsplänen Unterschiede zwischen Versuchsgruppen allein auf den Effekt einer Manipulation zurückgeführt werden können, da alle anderen Bedingungen zwischen den Gruppen konstant gehalten werden, gilt es in Feldstudien (wie z.B. Untersuchungen über die prognostische Validität von verschiedenen Prädiktoren für den Studienerfolg), eine Fülle möglicher Probleme in Versuchsplänen und in der statistischen Datenanalyse zu beachten, und so kausale Rückschlüsse nur mit großer Vorsicht zu ziehen. Im Folgenden wird daher detailliert darauf eingegangen, was es in Untersuchungen zu beachten gilt, die mit Survival Analysis analysiert werden sollen, insbesondere solchen mit zensierten Datensätzen. Die Darstellung richtet sich nach Empfehlungen von Singer & Willett (1991), auf die der interessierte Leser zur Vertiefung verwiesen sei.

Zum Versuchsdesign bei Survival Analysis

Stichprobenauswahl

Untersuchungen, die mit Survival Analysis (oder auch mit allen anderen statistischen Verfahren) analysiert werden sollen, können nur dann valide Resultate erbringen, wenn eine für eine bestimmte, wohldefinierte Zielpopulation repräsentative Stichprobe gewählt wird. Vor allem in Längsschnittstudien sollte daher zunächst die Zielpopulation definiert werden, um sicherzustellen, dass die interessierende Variable (im Falle einer Survival Analysis die Zeit), und insbesondere deren Streuung, generalisiert werden kann. Dies könnte vor allem dann infrage gestellt werden, wenn die Datensätze einer Stichprobe auf solche Individuen gekürzt würden, deren Zeitpunkte des interessierenden Ereignisses bekannt sind. Eine solche Strategie würde die mittlere Zeit bis zum interessierenden Ereignis unterschätzen. Ein anderes Problem ergibt sich in retrospektiven Studien: Bei Zusammenstellung einer Stichprobe stellt sich hier die Frage, ob es Individuen gibt, die eigentlich in die Stichprobe mit aufgenommen werden müssten, aber nicht erreichbar sind, weil sie das interessierende

Ereignis schon „hinter sich“ haben – in retrospektiven Studien über lebensbedrohliche Ereignisse also schon gestorben sind. Im oben genannten Beispiel einer retrospektiven Studie, in welcher der Zusammenhang zwischen Abiturnote und einem Kriterium für den Studienerfolg beleuchtet wird, wäre eine anfallende Stichprobe von allen Studierenden eines Semesters nach dem 12. Fachsemester nicht repräsentativ, da alle Studienabbrecher ausgeschlossen würden. Statistiker sprechen in diesem Fall vom Problem der *left-truncation.* Dieses ist nach Hutchison (1988, zitiert nach Singer & Willett, 1991) in der Literatur sehr oft vernachlässigt worden oder nicht vom Problem der *Links-Zensur,* auf die weiter unten eingegangen wird, unterschieden worden. Zusammenfassend gilt es daher, die Zielpopulation einer Untersuchung sorgfältig zu definieren und, wenn nicht anders möglich, auf eine erreichbare Subpopulation zu beschränken. Generalisiert man schließlich die Befunde auf die einer Untersuchung zugrunde liegende Zielpopulation, ist zudem darauf zu achten, ob Verallgemeinerungen kohortenabhängig sind und somit nur auf die Population einer bestimmten Kohorte, also beispielsweise auf bestimmte Abiturjahrgänge, zutreffen.

Definition des Zielereignisses

Wie bereits erwähnt, ist Survival Analysis zum einen an die Bedingung gebunden, dass jeder Proband zu jedem Zeitpunkt, an dem eine Messung erfolgen soll, einem und nur einem bestimmten Status zugeordnet werden kann; zum anderen findet ein Zielereignis statt, wenn sich der Status eines Probanden ändert. In diesem Kontext kann es insbesondere dann zu Schwierigkeiten kommen, wenn der Übergang von einem Status zum anderen kontinuierlich ist, oder Grenzen undeutlich sind. Daher ist es wichtig, alle möglichen Status so präzise wie möglich zu definieren. Singer & Willett (1991) empfehlen, Richtlinien zu erstellen, in denen a priori die jedem Status entsprechenden spezifischen Verhaltensweisen, Reaktionen oder Punktwerte aufgelistet werden, um die Zuordnung der Probanden zu einem Status zu erleichtern. Die Autoren verweisen hierbei auf die Literatur über Rückfallquoten bei Abhängigen, wo Variationen in der Definition des Rückfalls nach einem Entzug (sei es von Alkohol, Tabak, Nahrung oder Drogen) erklären können, warum verschiedene Wahrscheinlichkeiten für einen Rückfall zu gleichen Zeitpunkten nach einem Entzug gefunden wurden. Im Kontext der Studierendenauswahl durch die Hochschulen erscheint es für viele Studien beispielsweise plausibel, genau zu definieren, wann das Zielereignis „erfolgreicher Studienabschluss“ eingetreten ist – kann man bei einem mit der Gesamtnote „gut“ abgeschlossenem Diplom nach 20 Semestern oder bei einem „ausreichenden“ Diplom nach 9 Semestern immer noch von einem erfolgreichen Studienabschluss sprechen, d.h. liegt dann noch eine für die Universitäten wünschenswerte Passung zwischen den Anforderungen des Studiengangs und Merkmalen der Person vor?

Zeitlicher Beginn einer Studie und Eintrittszeitpunkt

Der „Startpunkt“, an dem die Beobachtung eines Probanden beginnt, ist von elementarer Bedeutung, wenn die interessierende abhängige Variable die Zeit bis zu einem Zielereignis ist und leitet sich direkt aus der Untersuchungsfrage ab. In experimentellen Untersuchungen wird nach Singer & Willett (1991) hierfür meist der Zeitpunkt der Randomisierung oder der Manipulation verwendet; in diesem Fall entsprechen sich der *zeitliche Beginn* einer Beobachtung im Leben eines Probanden und der *Eintrittszeitpunkt* in die Studie.

In Längsschnittuntersuchungen gestaltet sich die Wahl des zeitlichen Beginns einer Beobachtung schwieriger; so mag der am einfachsten zu ermittelnde Zeitpunkt nicht immer der beste sein. Möchte man etwa in einer entwicklungspsychologischen Studie wichtige Entwicklungsschritte im ersten Lebensjahr untersuchen, könnte der Zeitpunkt der Empfängnis als Startpunkt der Zeitmessung geeigneter sein als der Zeitpunkt der Geburt, um früh- und spätgeborene Kinder unterscheiden zu können. Probleme ergeben sich, wenn ein bestimmtes Ereignis, das als der zeitliche Beginn einer Beobachtung definiert wird, nicht für alle Probanden bekannt ist. Statistiker sprechen hierbei vom Problem der *Links-Zensur* (die vom Problem der *Rechts-Zensur,* wie zuvor beschrieben, zu unterscheiden ist). Singer & Willett zufolge umgehen die meisten Forscher dieses Problem, indem sie links-zensierte Datensätze von der Analyse ausklammern. Da Untersuchungen über die Prognostizierbarkeit von Studienerfolg höchstwahrscheinlich den Studienbeginn oder den Beginn eines Hauptstudiums als Eintrittszeitpunkt wählen und diese immer bekannt sind, bzw. recherchiert werden können, ist das Problem der Links-Zensur in diesem Kontext wenig relevant. Generell sollte jedoch immer darauf geachtet werden, den zeitlichen Beginn einer Beobachtung so zu wählen, dass das Problem der Links-Zensur eliminiert oder zumindest minimiert wird.

Neben der Wahl des geeigneten zeitlichen Beginns einer Beobachtung kann sich in Längs-schnittstudien ein weiteres Problem ergeben: im Gegensatz zu experimentellen Versuchs-anordnungen kann es passieren, dass sich der zeitliche Beginn und der *Eintrittszeitpunkt* in eine Untersuchung nicht entsprechen. Es sollte daher sorgfältig darüber nachgedacht werden, wie der „Eintritt in die Studie“ definiert ist. Ein *„später Eintritt“* (bzw. *verzögerter Eintritt*; *late entry*) beschreibt Situationen, in denen für einige oder alle Personen eine Verzögerung zwischen dem Zeitbeginn der Studie (spezifiziert durch die Untersuchungs-frage) und der Eintrittszeit (limitiert durch die verfügbaren Daten) besteht. Um auf das Beispiel der Evaluation eines fachspezifischen Studierfähigkeitstests für Studienbewerber zurückzukommen, würden sich für Quereinsteiger oder Studienortwechsler der zeitliche Beginn der Studie (einige Wochen vor Beginn des ersten Fachsemesters) und der Eintrittszeitpunkt (etwa nach dem ersten Fachsemester) nicht entsprechen; solche Probanden hätten einen verzögerten Eintritt in die Studie. Es stellt sich sodann die Frage, ob diese Probanden die gleiche Wahrscheinlichkeit für das Erreichen des Zielereignisses (z.B. ein erfolgreicher Abschluss der Diplom-Vorprüfung) haben wie diejenigen Personen,

die bereits unter Beobachtung stehen. Dies stellt ein geringfügiges Problem dar, wenn die Personen, die neu unter Beobachtung kommen, denjenigen *ähnlich* sind, die bereits beobachtet werden (zum Beispiel im Hinblick auf Abiturdurchschnittsnoten). In diesem Fall spricht man von *nicht-informativem späten Eintritt.* Kann dies jedoch nicht als gegeben angenommen werden (was oft der Fall ist), so muss jenem Umstand Rechnung getragen werden (zur Vertiefung, wie dies geschehen kann, siehe Bull & Spiegelhalter, 1997).

In Situationen mit einem späten Eintritt ist die Beobachtungsperiode (das Intervall zwischen der Eintrittszeit und dem Auftreten des Zielereignisses oder einer Zensierung) möglicherweise kürzer als die Überlebenszeit; sind dem hingegen der Zeitbeginn und der Eintrittszeitpunkt jedes Teilnehmers deckungsgleich, so sind Überlebenszeit und Beobachtungsperiode identisch.

Dauer einer Datenerhebung

Die zeitliche Länge einer längsschnittlichen Datenerhebung determiniert den Umfang der rechts-zensierten Daten: je länger alle Probanden verfolgt werden, desto geringer wird die Anzahl derer, die ein interessierendes Ereignis bis zum Ende einer Beobachtung (noch) nicht erreicht haben. Mit zunehmender Dauer der Datenerhebung in einer Längsschnittstudie steigen allerdings auch die Kosten und der Verlust von Probanden aufgrund anderer Gegebenheiten als einem interessierenden Ereignis, so dass es für jede Studie individuell gilt, die Vor- und Nachteile einer bestimmten zeitlichen Dauer abzuwägen. Zunächst muss daher eingeschätzt werden, wann ein bestimmtes Zielereignis eintritt. Singer & Willett (1991) empfehlen, ein Follow-Up dann anzusetzen, wenn davon ausgegangen werden kann, dass zu diesem Zeitpunkt mindestens die Hälfte der Probanden ein Zielereignis erreicht hat. Damit liegt genügend Information vor, den Median der „Überlebenszeit“ zu schätzen, und es kann davon ausgegangen werden, dass so statistisch signifikante Unterschiede zwischen Versuchsgruppen gefunden werden können (siehe *Tabelle 1*). Kann man also beispielsweise davon ausgehen, dass mehr als 50% der Studienanfänger einer bestimmten Fachrichtung die Diplom-Vorprüfung nach 4 Fachsemestern abgeschlossen haben, oder aber ihr Studium abgebrochen haben, könnte eine Follow-Up-Untersuchung zur Evaluation eines Studierfähigkeitstests 2 Jahre nach Studienbeginn stattfinden.

Follow-Up-Untersuchungen in prospektiven Studien

In prospektiven Studien kann die Datenerhebung entweder kontinuierlich, oder aber diskret erfolgen. In letzterem Fall gilt es – anhand der Fragestellung - zu entscheiden, ob eine einmalige Follow-Up-Untersuchung genügt, oder ob mehrere Messungen stattfinden sollen, und wenn ja, wie viele und zu welchen Zeitpunkten. In Untersuchungen, in denen der Übergang von einem Status in einen anderen kontinuierlich erfolgt, Messungen aber diskret in Intervallen erhoben werden, entscheidet der Zeitpunkt der Messungen und die Anzahl der

Intervalle über den Grad der Messgenauigkeit. Nun könnten mehrere Messungen in gleich bleibenden Intervallen durchgeführt werden. Um die Messgenauigkeit zu erhöhen, sollten in vielen Untersuchungen jedoch vermehrt Messungen zu solchen Zeitperioden gemacht werden, von denen erwartet wird, dass eine Vielzahl der Probanden von einem Status in einen anderen übergeht, und es können weniger Messungen zu solchen Zeiten gemacht werden, in denen sich das Zielereignis mit relativ geringer Wahrscheinlichkeit ereignet. Angenommen, in einer prospektiven Studie zur Evaluation der Vorhersagekraft eines Studierfähigkeitstests wählte man den Studienabbruch als Kriterium und würde die Beobachtungsperiode auf die Dauer von 9 Semestern ausdehnen, so sollten während der ersten Fachsemester Messungen in kürzeren Intervallen stattfinden, als gegen Ende der Regelstudienzeit, da dann gewöhnlich weniger Studienabbrüche vorliegen als im Grundstudium.

Drop-Outs in Längsschnittstudien

Je länger eine Untersuchung andauert, desto wahrscheinlicher kommt es zu Ausfällen von Probanden aufgrund anderer Gegebenheiten als dem interessierenden Ereignis. Im eben erwähnten Beispiel eines Längsschnitts zur Evaluation eines Studierfähigkeitstests erscheint es möglich, dass neben Studienabbrechern manche Studierende nicht an Follow-Up-Untersuchungen teilnehmen, weil sie die Bereitschaft an einer Teilnahme verloren haben, den Studienort gewechselt haben (aus Gründen, die nichts mit der Passung einer Universität und der Person zu tun haben), ein Praktikums- oder Auslandssemester absolvieren oder in der Zwischenzeit erkrankt oder gar gestorben sind. Während sich einige dieser Fälle nicht verhindern lassen, kann die Anzahl sogenannter Drop-Outs durch viele Maßnahmen gering gehalten werden, unter anderem indem den Probanden vorab klargemacht wird, warum nachfolgende Untersuchungen wichtig sind, oder indem Probanden im Falle eines Umzugs (bzw. im Falle eines Auslandssemesters etc.) gebeten werden, Kontaktadressen zu hinterlassen. Zudem können Belohnungen für die Teilnahme an jeder Untersuchung und insbesondere bei abgeschlossener Datenerhebung in Aussicht gestellt werden oder regelmäßig Newsletters via E-Mail verschickt werden. In jedem Fall sind die Daten solcher Ausfälle zensiert, und es gilt zu überprüfen, ob diese Ausfälle wirklich *nicht-informativ* sind, oder ob sich solche Probanden systematisch von denen unterscheiden, die bis zum Ende einer Untersuchung teilnehmen (in diesem Fall läge eine selektiver Drop-Out vor, der in diesem Zusammenhang als *informative Zensur* bezeichnet würde und unter der fälschlichen Annahme einer nicht-informativen Zensur möglicherweise in einer Verzerrung der Schätzungen und unzulässigen statistischen Schlussfolgerungen resultieren würde). Nach Singer & Willett (1991) eignen sich zwei Strategien besonders, mit solchen Daten umzugehen: Entweder wird für jeden ausgefallenen Probanden angenommen, dass sich das Ereignis bis zur Zensur (also dem Zeitpunkt, an dem der Status eines Probanden zum letzten Mal beobachtet wurde) nicht ereignet hat, oder aber man

nimmt an, dass sich der Status genau danach verändert hat. Welche Alternative zu bevorzugen ist, hängt vom Zweck der Untersuchung ab; letztgenannte eignet sich in manchen Forschungsbereichen besser, wie zum Beispiel im Kontext der Forschung über Rückfälle nach einem Entzug, wo Forscher davon ausgehen, dass Probanden in Kontakt blieben, wenn sie weiterhin „clean" wären.

Mögliche Fehlerquellen in retrospektiven Studien

Generell sind prospektive Studien retrospektiven vorzuziehen. Dennoch können Fragestellungen über sehr seltene Ereignisse vorliegen, die eine längsschnittliche Beobachtung unmöglich machen oder zumindest aus ökonomischen Gründen ausschließen. So interviewen viele Forscher ihre Probanden retrospektiv und fragen diese, ob sich eine bestimmte Gegebenheit in ihrem Leben ereignet hat, und wenn ja, wann. Abgesehen vom Problem der left-truncation kann es dann zu folgenden Fehlern kommen: Probanden haben völlig vergessen, dass sich eine interessierende Gegebenheit überhaupt ereignet hat, was zu einer Unterschätzung des Auftretens eines Ereignisses führt. Denkbar ist auch, dass Probanden den Zeitpunkt eines Ereignisses falsch einschätzen, insbesondere der Meinung sind, etwas habe sich kürzlicher ereignet, als es tatsächlich der Fall war („*telescoping*"), oder solche Zeitpunkte aufrunden („vor zweieinhalb Jahren"), was ebenfalls zu einer Verzerrung der Überlebensfunktion führen kann. So sollte individuell für jede spezifische Fragestellung geprüft werden, ob solche Fehler in einer geplanten Untersuchung vorkommen können. Gegebenenfalls finden sich im Artikel von Singer & Willett (1991) verschiedene Maßnahmen, durch die solche Fehler gering gehalten werden können, wie beispielsweise Erinnerungshilfen durch eigens konstruierte Items.

Zielereignisse, die sich wiederholen können

Studien, die sich mit der verlängerten Lebenserwartung nach einem medizinischen Eingriff oder mit der Zeit bis zum erfolgreichen Abschluss eines Studiums befassen, erheben die Zeit bis zu einem Zielereignis, das einmalig im Leben eines Probanden ist. Untersucht man aber solche Ereignisse, die sich wiederholen können (z.B. depressive Episoden, Rückfall nach einem Entzug, Verbrechen, usw.) sollte erhoben werden, ob ein solches Ereignis zum ersten, zweiten, …, x-ten Mal geschieht, da vorangehende (wenn auch „nicht erfolgreiche") Behandlungen die Wahrscheinlichkeit für den Erfolg einer nachfolgenden Behandlung erhöhen.

Stichprobengröße

Auch bei den Methoden der Survival Analysis ist es wichtig, Effektstärken zu testen und so Aussagen über die Wahrscheinlichkeit für bestimmte Effekte in der Zielpopulation einer Untersuchung zu ermöglichen. Im Folgenden soll abschließend am Beispiel einfacher

Zweigruppen-Vergleiche erläutert werden, wie sich die minimal erforderliche Größe einer Stichprobe für Untersuchungen beliebiger zeitlicher Dimensionen berechnet.

Um über den Umfang einer geplanten Stichprobe zu entscheiden, müssen zunächst statistische Hypothesen formuliert werden sowie die erwünschten Irrtumswahrscheinlichkeiten und die minimale erwünschte Effektgröße bestimmt werden. Für die Durchführung einer Survival Analysis müssen Forscher zudem die Verteilung der Hazard-Funktion (die im Folgenden erläutert wird) und die Länge eines Follow-Ups benennen.

Tabelle 1 veranschaulicht eine Übersicht über die jeweils von der Länge des Follow-Up abhängige minimale Stichprobengröße, mit der bestimmte Effektgrößen erreicht werden können. Die Werte belaufen sich auf eine statistische Stärke von jeweils .80 für einen einfachen Zweigruppen-Vergleich mit einer zweiseitigen Irrtumswahrscheinlichkeit von 0.05. Die Effektgrößen berechnen sich in diesem Beispiel durch das Verhältnis der medianen Lebensdauer in beiden Gruppen, d.h. bei einer Effektgröße von 1.25 ist die mediane Lebensdauer in der einen Gruppe um 25% höher als in der anderen Gruppe. Um zu entscheiden, welche Effektgröße mindestens erreicht werden sollte, können Forscher recherchieren, ob in der Fachliteratur bereits Befunde über Effektgrößen in ähnlichen Vergleichen vorliegen. Liegen solche Informationen nicht vor, empfehlen Schoenfeld & Richter (1982, zitiert nach Singer & Willett, 1991), von einer Effektgröße von 1.50 auszugehen, da eine um 50% erhöhte Lebensdauer klinisch bedeutsam sei.

Tab. 1: Minimaler Stichprobenumfang zur Berechnung von Unterschieden zwischen Gruppen in Längsschnittstudien mit zeitlich beliebigen Dimensionen der Follow-Up-Messungen

	Follow-Up-Periode				
Effektgröße	0.5	1.0	1.5	2.0	2.5
1.25	>2162	1260	976	840	766
1.50	654	382	296	254	232
1.75	344	200	156	134	122
2.00	224	130	102	88	80

Die Länge einer Follow-Up-Periode berechnet sich als das arithmetische Mittel der erwarteten medianen Lebensdauern in beiden Gruppen, d.h. bei einer Follow-Up-Messung von 0.5 erfolgt eine Nachuntersuchung nach der halben durchschnittlichen erwarteten medianen Lebensdauer. Dabei kann die tatsächliche Länge einer späteren Untersuchung beträchtlich variieren; die Werte können sowohl für Settings mit Nachuntersuchungen innerhalb von Minuten, Tagen wie auch Jahren verwendet werden. Wie in Tabelle 1

deutlich wird, sind kleine Effektgrößen schwieriger zu entdecken als größere, für die ein kleinerer Stichprobenumfang benötigt wird. Ebenfalls hängt der Stichprobenumfang von der Länge der Follow-Up-Periode ab in dem Sinne, dass sich mit zunehmender Länge die Anzahl der minimal benötigten Probanden verkleinert.

Zusammenfassend können durch eine längere Beobachtung von Probanden leichter statistisch signifikante Befunde erreicht werden sowie der Umfang einer Stichgröße reduziert werden. Daher empfiehlt es sich, Probanden in einer Längsschnittstudie mindestens für die Zeit der durchschnittlich zu erwartenden medianen Lebensdauer zu beobachten.

Werden nun all diese Hinweise zur Anwendung der Survival Analysis bereits bei der Planung von jeglichen Untersuchungen zur Prognostizierbarkeit von Studienerfolg bedacht, kann man - sofern sinnvolle Fragestellungen vorliegen - davon ausgehen, valide Datensätze zu erhalten, mit denen Rückschlüsse auf die Vorhersagegüte verschiedener Prädiktoren möglich sind. Im Anschluss daran und unter Beachtung der rechtlichen Rahmenbedingungen können sodann solche Prädiktoren eventuell bei der Entwicklung eines Auswahlverfahrens von Studienbewerbern miteinbezogen werden. Der zweite Teil dieser Arbeit soll sich nun damit befassen, wie solche Stichprobendaten statistisch zu analysieren sind.

[Literaturverzeichnis folgt am Ende des Beitrags von Dorothea Mildner]

Survival- und Hazardfunktionen

Dorothea Mildner

Einführung

Durch den Gebrauch der Methoden der Survival Analysis ist der Untersucher in der Lage, Auftretensmuster von Ereignissen zu erkennen und zu beschreiben, diese Muster zwischen Gruppen zu vergleichen und statistische Modelle vom Risiko des Auftretens von kritischen Begebenheiten über die Zeit zu formulieren.

Ausgehend von seinen Wurzeln in den Untersuchungen der menschlichen Lebenszeit, wo das Zielereignis in der Regel der Tod ist, ist der Terminus „Überlebensanalyse“ eher negativ behaftet. Aber hinter der Terminologie liegt ein eine starke Methode, welche die Daten von *allen* Beobachtungen nutzt, zensierte und unzensierte Fälle gleichermaßen. Betont werden muss an dieser Stelle, dass der Begriff „Überleben“ keinesfalls nur mit „Tod“ oder Ähnlichem in Verbindung gebracht werden darf. Eine „Lebensdauer“ (*survival time*) kann gleichsam die Zeit bis zu der Entlassung aus einem Krankenhaus, bis zur Heirat oder bis zu einem Aufmerksamkeitswechsel im Klassenraum umfassen.

Die Survival Analysis wird oftmals benutzt in industriellen Haltbarkeits-Testungen („Lebensdauer“ von verschiedenen Produkten, z.B. von Glühbirnen), und sie wird häufig verwendet in klinischen und epidemiologischen Follow-Up-Studien. Generell haben sich in den letzten zwei Jahrzehnten viele neue Anwendungsfelder für die Methoden der Survival Analysis erschlossen, die zum Teil weit entfernt von den ursprünglichen (siehe oben) liegen.

Bei der Auswahl von Studienanfängern erhalten deutsche Hochschulen zukünftig einen größeren Spielraum. Dies fördert Profilbildung zwischen Universitäten und den Wettbewerb um besonders qualifizierte Studierende. Nach einer Änderung des Hochschulrahmengesetzes (HRG) werden ab dem Wintersemester 2005/2006 60% der Studienplätze in den zentralen Numerus Clausus – Fächern direkt durch die Hochschulen vergeben. Das eröffnet Möglichkeiten für Hochschulen und Studierende: Die Hochschulen können sich die Studierenden aussuchen, für die die Studienangebote die passenden sind. Andererseits werden die Bewerber bei der Wahl des Studiengangs unterstützt, indem ihre Eignung bzw. Nichteignung für den speziellen Studiengang schon *vor* der Aufnahme des Studiums prognostiziert wird. Allerdings sehen sich die Hochschulen auch mit der Aufgabe konfrontiert, ein hinreichend objektives, zuverlässiges, valides sowie faires und ökonomisches Bewerberauswahlverfahren zu entwickeln. *Rindermann & Oubaid (1999)* schlagen als adäquate Lösung ein mehrstufiges flexibles Auswahlmodell (das *Abitur-Test-Interview-Modell*; *ATIM*) vor, unter Einbezug von Abiturdurchschnittsnoten, Fachnoten,

Eignungstests und Aufnahmegesprächen, das fachspezifische Gewichtungen der einzelnen Prädiktoren ermöglicht.

Im Zusammenhang mit den Kriterien, Verfahren und der Prognostizierbarkeit des Studienerfolges, ist der Einsatz der Methoden der Survival Analysis indiziert und notwendig.

Nach Rindermann und Oubaid (1999) kann der Studienerfolg an sechs Kriterien festgemacht werden, von denen hier zwei exemplarisch herausgegriffen werden sollen:

1. Der *Studienabschluss* stellt ein basales Erfolgskriterium dar, kann doch bei *Studienabbruch* kaum von einem erfolgreichen Studium gesprochen werden (Gold & Kloft, 1991; nach Rindermann & Oubaid, 1999). Die Fälle, in denen das Studium aufgrund einer erfolgreichen Berufskarriere in einem dem Studienfach zuordenbaren Bereich aufgegeben wurde, dürften selten sein. Der Studienabbruch muss von Universitätswechsel und Studienfachwechsel unterschieden werden.
2. Die *Studiendauer* wurde in den letzten Jahren aufgrund der oft kritisierten, von der Regelstudiendauer bedeutend abweichenden Studienzeiten häufiger berücksichtigt (Daniel, 1996; Giesen & Gold, 1996; Reissert, 1991; nach Rindermann & Oubaid, 1999). Ein Studium gilt dann als erfolgreich, wenn in kurzer Zeit ein qualifizierter Abschluss erreicht wurde. Die Erfassung dieser Variable wird durch die Defizite der Hochschulstatistik erschwert. Beispielsweise verzerren Quereinsteiger, Fach- und Ortswechsler oder wegen finanzieller Vorteile Immatrikulierte die Studienstatistiken (Wagemann, 1987; nach Rindermann & Oubaid, 1999).

Besonders im Hinblick auf diese zwei genannten Kriterien kann der Gebrauch der Methoden der Überlebensanalyse angezeigt und hilfreich sein. Beispielsweise kann es von Interesse sein, die Wahrscheinlichkeit eines Studienabbruches in Abhängigkeit von der Zeit zu prognostizieren. Es kann etwa der Blick darauf gerichtet werden, wann (zeitlich gesehen) erhöhte Risikoperioden bestehen, ein Studium abzubrechen oder auch verschiedene Subgruppen von Studierenden (z.B. im Hinblick auf sozioökonomische Unterschiede oder Schulabschlüsse aus verschiedenen Bundesländern) oder verschiedene Fächergruppen in ihren spezifischen Risikoprofilen zu vergleichen.

Der interessierte Leser sei für eine Vertiefung der verschiedenen Methoden der Survival Analysis an dieser Stelle auf den Überblicksartikel von Bull & Spiegelhalter (1997) und das Buch von Harrell (2001) verwiesen. Singer & Willett (1991) geben zusätzlich praktische Design-Hinweise für Längsschnittstudien, in welchen die Zeit bis zum Eintritt eines bestimmten Ereignisses von Relevanz ist.

Survival- und Hazardfunktionen

Eine einfache *Deskription* von Resultaten aus Längsschnittuntersuchungen zu einem festgesetzten Zeitpunkt ist in zweierlei Hinsicht limitiert. Erstens bezieht diese nur mit ein, *ob* ein Ereignis zu einer bestimmten Zeit aufgetreten ist oder nicht; und Zweitens beinhaltet sie nur Personen, die während der *gesamten* Zeit verfolgt wurden, respektive hätten verfolgt werden können. Im Gegensatz hierzu nutzt die Survival Analysis die Information der gesamten Follow-up-Periode und *alle* Personen können Informationen beitragen während ihrer Zeit unter Beobachtung.

Innerhalb einer deskriptiven Analyse von Daten aus solchen Untersuchungen, kann man ein beobachtetes Verhältnis (*proportion*) p = (Anzahl derjenigen, die das Zielereignis erlebt haben / Gesamtanzahl der Personen) erfassen, das Werte zwischen 0 und 1 annehmen kann. Dieses Verhältnis kann als Risiko ausgedrückt werden, welches sich berechnet durch $p/(1 - p)$ = (die Anzahl derjenigen, die das Zielereignis erlebt haben / die Anzahl der „Überlebenden“) und Werte von 0 bis Unendlich annehmen kann. Das 95%ige (oder natürlich auch jedes andere; hier exemplarisch demonstriert für 95%) Konfidenzintervall für den wahren zugrundeliegenden Wert – zum Beispiel die wahre Sterberate oder die wahre Studienabbruchsrate – kann ebenfalls bestimmt werden. Es kann unter Benutzung der Standard-Normal-Approximation berechnet werden nach $p \pm 1{,}96 \cdot \sqrt{[p(1-p)/n]}$ (für Details siehe Bull & Spiegelhalter, 1997; p. 1047).

Generalisierungen, die aus solchen simplen Analysen folgen, hängen stark von der Repräsentativität der Stichprobe der beobachteten Personen im Hinblick auf die interessierende Gesamtpopulation ab.

Generell ist es das Ziel einer Überlebensanalyse, die verfügbaren Informationen zu nutzen, um Schätzungen über das Überleben (beziehungsweise darüber, frei von dem kritischen Ereignis zu sein) zu verschiedenen Zeiten abzugeben. Diese Beziehung wird als *Überlebensfunktion* (*survival function*; siehe Gl. 1) ausgedrückt. Eine graphische Darstellung der Überlebensfunktion bereitet die reizvollste Zusammenfassung der zeitbezogenen Informationen einer Datenmenge.

In der Survival Analysis wird T verwendet, um die Reaktionsvariable zu bezeichnen, da die Reaktion üblicherweise die Zeit bis zum Eintreten eines bestimmten Ereignisses ist. Anstatt ein statistisches Modell für die Reaktion T in Begriffen der erwarteten Überlebenszeit zu definieren, ist es vorteilhaft, es in Begriffen der *Überlebensfunktion* (*survival function*) $S(t)$ zu definieren, die gegeben ist durch:

$$S(t) = \text{Prob}(T>t) = 1 - F(t), \qquad \text{(Gl. 1)}$$

wobei $F(t)$ die kumulierte Verteilungsfunktion für T ist. Falls das Zielereignis zum Beispiel der Studienabbruch sein sollte, ist $S(t)$ die Wahrscheinlichkeit, dass der Abbruch *nach* Zeit t eintritt, das heißt, die Wahrscheinlichkeit, dass die Person wenigstens bis zu dem Zeitpunkt

t „überlebt“ (weiterstudiert). *S(t)* ist bei $t = 0$ immer 1; alle Individuen überleben zumindest Zeitpunkt Null. Die Überlebensfunktion verläuft mit fortschreitender Zeit nichtsteigend. Sind Personen einer Studie zu einem Zeitpunkt besonders hoch gefährdet, so fällt die Überlebensfunktion in diesem Bereich stark ab.

Die *kumulierte Hazardfunktion* wird mit $\Lambda(t)$ bezeichnet. Sie beschreibt das kumulierte Risiko des Erlebens des Zielereignisses bis zum Zeitpunkt *t*, und ist (wie beispielsweise bei Harrell, 2001 gezeigt wird) gleich dem negativen Logarithmus der Überlebensfunktion (siehe Gl. 3 und Gl. 4). $\Lambda(t)$ ist nichtfallend mit fortschreitender Zeit; das heißt, das kumulierte Risiko steigt oder bleibt gleich.

Eine weitere wichtige Funktion ist die *Hazardfunktion*, *h(t)*, die auch *integrierte Ausfallrate* genannt wird. Die Ausfallrate zum Zeitpunkt *t* bezieht sich auf die Wahrscheinlichkeit, dass ein Ereignis in einem schmalen Intervall um *t* herum auftauchen wird, unter der Annahme, dass das Ereignis vor Zeit *t* noch nicht eingetreten ist. Wenn man die Ereignisrate zu einem bestimmten Zeitpunkt (unter der Bedingung, dass die Begebenheit zu diesem Zeitpunkt noch nicht aufgetreten ist) betrachtet, kann man Aufschluss gewinnen über die Mechanismen und Einflüsse des Risikos über die Zeit. Die Hazardfunktion erlaubt es, Phasen erhöhten Risikos leichter zu identifizieren, als wenn nach plötzlichen Steigungsänderungen in *S(t)* oder $\Lambda(t)$ gesucht wird. Das Integral der Hazardfunktion *h(t)* bildet die kumulierte Hazardfunktion $\Lambda(t)$ und diese ist gleich dem negativen natürlichen Logarithmus der Überlebensfunktion *S(t)* (vgl. Gl. 2).

$$\int_0^t h(t) = -\log S(t) \qquad \text{(Gl. 2)}$$

$$\Lambda(t) = -\log S(t), \qquad \text{(Gl. 3)}$$

$$\text{oder } S(t) = \exp[-\Lambda(t)]. \qquad \text{(Gl. 4)}$$

Eine dieser drei Funktionen zu kennen, erlaubt folglich dem Untersucher, die anderen beiden abzuleiten. Die drei Funktionen sind unterschiedliche Wege, die gleiche Verteilung zu beschreiben. *Abbildung 2* zeigt eine graphische Gegenüberstellung der Funktionen. Angesichts der glatten Überlebensfunktion in A. und der ebenfalls glatten Hazardfunktion in C. lässt sich schließen, dass der Schätzung der Survivalfunktion ein parametrisches Verfahren zugrunde liegen muss. Auf diesen Sachverhalt wird an späterer Stelle vertiefend eingegangen werden.

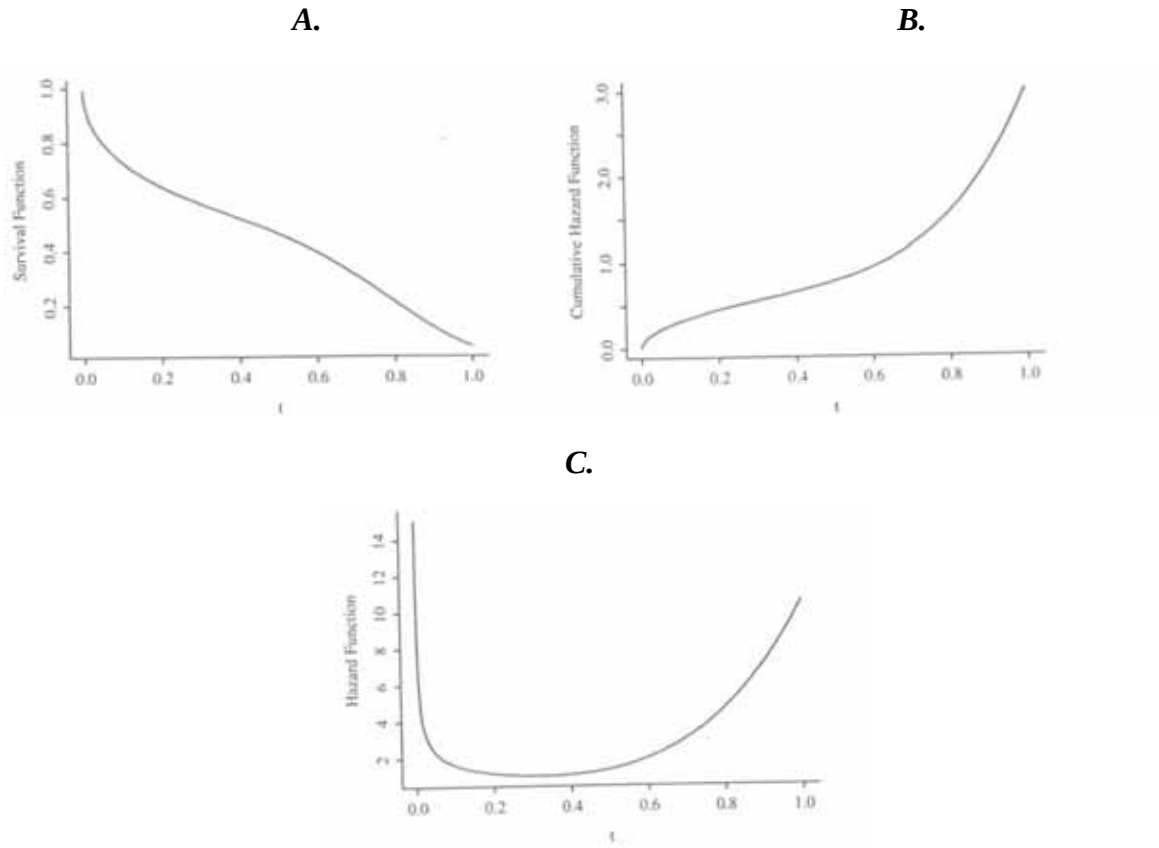


Abbildung 2: Die Abbildung zeigt korrespondierende A. Survival-, B. kumulierte Hazard- und C. Hazardfunktionen (entnommen aus Harrell, 2001; p. 393-394). Auf der Abszisse findet sich jeweils die Zeit abgetragen; auf der Ordinate findet sich bei A. die Wahrscheinlichkeit, einen Zeitpunkt t zu überleben, bei B. findet sich das kumulierte Risiko, und bei C. die Ausfallrate.

Überlebens- oder Zuverlässigkeitsfunktionen (survival functions)

Die Überlebensanalyse beginnt mit der *Überlebens- oder Zuverlässigkeitsfunktion*. Untersucht man beispielsweise die Effizienz eines Raucherentzugsprogramms (Daten aus Stevens & Hollis, 1989; zitiert nach Singer & Willett, 1991), repräsentiert die Überlebensfunktion die Wahrscheinlichkeit, dass ein zufällig ausgewählter Ex-Raucher abstinent bleibt versus Zeit (s. Abb. 3).

Zum Beginn der Studie (hier auch gleichzeitig der Beginn der „Zeit“) ist die Überlebenswahrscheinlichkeit 1,00. Mit der Zeit und mit den zunehmenden Rückfällen fällt die Überlebensfunktion gegen 0. In dieser Studie sind 82% der Teilnehmer der Studie abstinent („überleben“) mehr als vier Wochen, 66% sind mehr als acht Wochen abstinent und 60% rauchen länger als 12 Wochen nicht. Nach 50 Wochen, dem Ende der Datenerhebung, bleiben 38% rauchfrei. Diese Personen haben zensierte Rückfallzeiten, entweder, weil sie niemals rückfällig werden oder sie es werden, nachdem die

Datensammlung endet. Aufgrund der Zensierung erreichen Überlebensfunktionen selten 0. Diese Funktion enthält also sowohl unzensierte Daten von denjenigen Personen, welche innerhalb der 50 Wochen wieder rauchten, als auch zensierte Daten von jenen Teilnehmern, die bis zu diesem Zeitpunkt nicht rückfällig wurden.

Die Überlebensfunktion der Stichprobe liefert Informationen, um folgende deskriptive Frage zu beantworten: Wie viele Wochen vergehen, bevor der durchschnittliche Raucher rückfällig wird?

Erreicht die Funktion .50, bedeutet das, dass die Hälfte (50%) der Ex-Raucher rückfällig geworden sind und die andere Hälfte nicht. In diesem Beispiel liegt der Punkt, welcher anzeigt, wie viel Zeit vergeht, bevor die Hälfte der Stichprobe die Zielbegebenheit (den Rückfall) erlebt, bei etwa 16 Wochen.

Alle Überlebensfunktionen haben eine ähnliche Form wie die in Abb. 3 gezeigte: eine negativ beschleunigte (Löschungs-) Kurve, eine monoton nichtsteigende Funktion der Zeit. Diese Generalisierung wurde von Hunt und Kollegen schon weit vor der Verbreitung der modernen Überlebensanalyse hervorgehoben (Hunt et al., 1971; Hunt & General, 1973; nach Singer & Willett, 1991). Nachdem sie ähnliche Überlebensmuster in nahezu 100 Studien zum Raucher-, Heroin- und Alkoholentzug fanden, sagten Hunt et al. (1971) die Nützlichkeit der Überlebensfunktion vorher, indem sie schrieben, dass sie „hofften, die Steigungs- (resp. Abfall-) Unterschiede zwischen individuellen Kurven als ein differentielles Kriterium zu nutzen, um verschiedene Treatmenttechniken zu evaluieren“ (zitiert nach Singer & Willett, 1991; p.270).

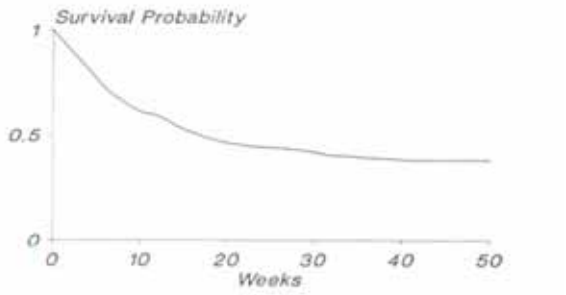


Abbildung 3: Überlebenskurve (nach den Daten von Stevens & Hollis, 1989; entnommen aus: Singer & Willett, 1991, p.270). Auf der Abszisse ist die Zeit abgetragen, auf der Ordinate die Wahrscheinlichkeit, einen Zeitpunkt t zu „überleben“.

Das geschilderte Vorgehen lässt sich mühelos auch auf ein Beispiel aus dem universitären Kontext beziehen. Es kann den Untersucher etwa interessieren, wie viele der in einem Studiengang immatrikulierten Studierenden bis zu einem festgesetzten Zeitpunkt ihr Studium abbrechen. Die Überlebensfunktion repräsentiert in diesem Fall die Wahrscheinlichkeit, dass ein zufällig ausgewählter Studierender sein Studium fortführt versus Zeit. Zu

Beginn der Untersuchung (angenommen, dies sei die Immatrikulation), der auch gleichzeitig den Eintrittszeitpunkt der Personen in die Studie darstellt, ist die Überlebenswahrscheinlichkeit 1. Alle Studierenden „überleben" zumindest den Zeitpunkt der Immatrikulation ($t = 0$). Würde man nun am Ende eines jeden Semesters (hier besteht also eine Intervall-Zensierung) erfassen, wie viele der betrachteten Studierenden noch studieren, so ließen sich Aussagen treffen, wie „78% der Studierenden studieren auch nach einem Semester noch; 61% studieren länger als zwei Semester" usf. Die Überlebensfunktion verläuft mit fortschreitender Zeit nichtsteigend. Ist der durchschnittliche Studierende zu einem Zeitpunkt (beispielsweise in dem Semester vor Ablegen einer Zwischenprüfung) besonders hoch gefährdet, sein Studium abzubrechen, so fällt die Überlebensfunktion in diesem Bereich stärker ab. Die Daten können von rechts zensiert sein, durch einen vom Untersucher festgelegten Zeitpunkt des Endes der Datenerhebung (etwa das zweite Semester nach Ablegen einer Zwischenprüfung o.ä.). Studierende können gleichsam aus anderen Gründen dem Follow-Up verloren gehen (siehe oben).

Non-parametrische Überlebensfunktionen

Zuerst soll illustriert werden, wie die Zensierung eingebettet werden kann in das sogenannte *Kaplan-Meier-(K-M-)Verfahren*. Dieses ist bekannt als ein non-parametrischer Weg, eine Überlebensfunktion zu schätzen, da es keine Annahmen über die Form der zugrundeliegenden Überlebenskurve macht (es geht nicht davon aus, dass die Überlebenskurve mathematisch zusammengefasst werden kann durch eine begrenzte Anzahl von Parametern). Diese Art der Schätzung kann sehr sinnvoll sein, denn die wahre Form der Überlebensverteilung ist nur selten bekannt. In vielen Analysen mag eine solche Schätzung der letzte Schritt sein, wohingegen in anderen Analysen dieses Vorgehen helfen kann, ein statistisches Modell für eine tiefergehende Analyse auszuwählen. Ein naiv-explorativer Einsatz dieses Verfahrens ist denkbar. Die K-M-Methode schätzt das augenblickliche Risiko des Zielereignisses zu jeder bestimmten Zeit als das Verhältnis der Anzahl derjeniger, die das kritische Ereignis zu dieser Zeit erlebt haben und der Anzahl derer, in der gegenwärtigen Risikogruppe. Die Risikogruppe setzt sich aus denjenigen Personen zusammen, die zurzeit das Risiko haben, das kritische Ereignis zu erleben.

Tabelle 2 soll diesen Sachverhalt an einem Datenbeispiel verdeutlichen (modifiziert nach Bull & Spiegelhalter, 1997; p. 1049). Ausgegangen wird von einem 30 Personen umfassenden Datenset, aus dem hier ein exemplarischer Ausschnitt dargestellt ist. Zum Zeitpunkt des ersten Studienabbruchs (von Person 1) vier Tage nach Aufnahme des Studiums, sind 30 Personen in der Risikogruppe und folglich beträgt das Risiko des Studienabbruchs am vierten Tag nach Beginn des Studiums 1/30=0,033. Das bedeutet, dass die Chance, *mehr* als vier Tage zu „überleben" (weiter zu studieren) geschätzt werden kann als 1-1/30=0,967, mit einem 95%igen Konfidenzintervall zwischen 0,79 und 0,99 (Berechnung siehe oben). Vierzehn Tage nach dem Studienbeginn bricht Person 15 ab, und

die Risikogruppe umfasst nun noch 29 Individuen. Die Chance, den vierzehnten Studientag zu „überleben“ wird geschätzt als 1-1/29=0,965, und die kumulative Wahrscheinlichkeit, mehr als 14 Tage zu studieren, beträgt 0,967 x 0,965=0,933 mit einem 95%igen Konfidenzintervall zwischen 0,76 und 0,98. Diese Berechnungen können in gleicher Art und Weise fortgeführt werden (vgl. Gl. 5). Die Ergebnisse in der Tabelle wurden auf zwei Nachkommastellen gerundet, um die generelle Widergabe solcher Resultate als „Prozent Überleben“ zu reflektieren. Werden die Wahrscheinlichkeiten mit 100 multipliziert, erhält man die prozentuale Wahrscheinlichkeit, über eine bestimmte Zeit hinaus zu überleben.

Tabelle 2: Datenbeispiel (modifiziert nach Bull & Spiegelhalter, 1997; p. 1049). In der Tabelle finden sich die Überlebenszeiten der Personen in Tagen, die Größe der Risikogruppe, der Kaplan-Meier-Schätzer des Überlebens und das zugehörige 95%ige Konfidenzintervall[3].

Person	**Überlebenszeit (in Tagen)**	**Risikogruppe (N)**	**K-M Überlebensschätzer**	**95% KI des Überlebensschätzers**
1	4	30	0,97	0,79 – 0,99
15	14	29	0,93	0,76 – 0,98
21	77	28	0,90	0,72 – 0,97
19	88	27	0,87	0,68 – 0,95
3	117	24	0,83	0,64 – 0,93
2	121	22	0,79	0,60 – 0,90
30	142	21	0,76	0,55 – 0,88
29	193	20	0,72	0,51 – 0,85
14	247	19	0,68	0,47 – 0,82
17	275	18	0,64	0,44 – 0,79
12	393	17	0,60	0,40 – 0,76
28	1100	15	0,56	0,36 – 0,73
16	1791	12	0,52	0,31 – 0,69
24	2982	10	0,47	0,26 – 0,64
11	3098	7	0,40	0,20 – 0,60

Abbildung 4 zeigt die (entsprechend dem Datensatz aus Tab. 2) geschätzte Überlebenskurve in dem aus dem Kaplan-Meier-Verfahren typischerweise resultierenden treppenstufenförmigen Verlauf.

[3] In der Tabelle 2 ist zu erkennen, dass die Risikogruppe bei dem Ausscheiden einer Person nicht immer gleichmäßig auch um eine Person schrumpft. Es kann hierbei davon ausgegangen werden, dass Personen aus irgendeinem Grund außer dem Erleben des Zielereignisses (Studienabbruch) dem Follow-Up verloren gegangen sind (z.B. Umzug, Studienfachwechsel o.ä.).

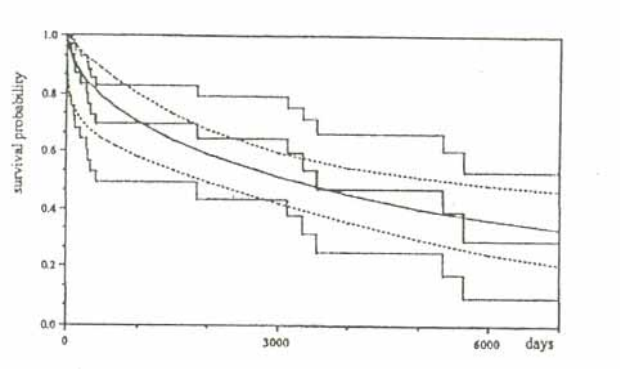


Abbildung 4: Mit dem Datensatz aus Tabelle 2 korrespondierende Kaplan-Meier-Schätzungen des Überlebens mit 95%igem Konfidenzintervall (obere und untere Grenzen aufgezeigt). Vergleichend dargestellt ist in der Abbildung auch die Weibull-Schätzung (siehe unten) samt ihren Konfidenzgrenzen (entnommen aus Bull & Spiegelhalter, 1997; p. 1050).

Mathematisch ausgedrückt, wird angenommen, dass es r_k Personen in der Risikogruppe zum Zeitpunkt der *k*ten bestimmten Zeit t_k gibt, und dass es zu dieser Zeit f_k Ereignisse (zum Beispiel Studienabbrüche) gegeben hat. Die geschätzte Überlebenswahrscheinlichkeit bis zu der Zeit t_K wird berechnet durch

$$p_K = \left(1 - \frac{f_1}{r_1}\right) \cdot \left(1 - \frac{f_2}{r_2}\right) \cdot .... \cdot \left(1 - \frac{f_K}{r_K}\right). \qquad \text{(Gl. 5)}$$

Die Kaplan-Meier-Schätzer sind in den meisten gängigen statistischen Computerpaketen implementiert, allerdings ist darin selten eine Einstellung für den späten Eintritt vorgesehen (für ausführlichere Information siehe Bull & Spiegelhalter, 1997).

Bull & Spiegelhalter (1997) vergleichen zwei nach dem Kaplan-Meier-Verfahren geschätzte Überlebensfunktionen miteinander, wovon eine das generelle geschätzte Überleben von Personen darstellt, die von der Geburt an beobachtet wurden und die andere Kurve nur Personen miteinbezieht, die ab dem Auftreten bestimmter Merkmale (hier der „Ausbruch“ einer Krankheit) unter Beobachtung stehen. Sie stellen fest, dass die Funktion, die einen Eintritt ab der Geburt annimmt, einen viel optimistischeren Verlauf darstellt, als die andere Funktion. Sie zeigen, dass der Verlauf der Kurven reflektiert, dass die Unterschiede zwischen den Funktionen nur durch die Größe der Risikogruppe bedingt ist. Die Wahl der Risikogruppe determiniert die Größe des Nenners in der Berechnung, und ein großer Nenner wird immer das gegenwärtige Risiko schmälern.

Parametrische Überlebensfunktionen

Non-parametrische Techniken benutzen die zur Verfügung stehenden Daten, um die Überlebensfunktion direkt „abzuzeichnen“. Diese Methodologie wird mit jedem Muster des Überlebens kompatibel sein, aber die Überlebensfunktion schreitet mit abfallenden Stufen fort, was nicht die übliche Wahrnehmung einer zugrundeliegenden Kontinuität widerspiegelt. *Parametrische* Überlebensfunktionen reflektieren beides, die Daten *und* einige Annahmen darüber, einschließlich einer zugrundeliegenden Kontinuität (zur Vertiefung siehe Harrell, 2001). Diese Annahmen sind eingebettet in Parameter, die selbst von den Daten her geschätzt werden; die resultierende Überlebensfunktion ist somit eine mathematische Gleichung, welche eine glatte Kurve beschreibt. Einfache parametrische Funktionen beinhalten die *Expotential-* (in welcher ein einziger Parameter das Überleben charakterisiert, das als konstant angenommen wird) und die *Weibullfunktion* (mit zwei Parametern, die es beispielsweise einer Todesrate erlauben, mit der Zeit entweder zu steigen oder zu sinken).

Die Abbildung 4 zeigt eine Weibullkurve für das „Überleben“, zusammen mit ihrem 95%-Konfidenzintervall. Die vergleichbare Kaplan-Meier-Kurve ist ebenfalls mit ihrem Konfidenzintervall aufgezeigt. Es lässt sich erkennen, dass das Konfidenzintervall für die parametrische Kurve *enger* ist, als für die non-parametrische Kurve. Diese zusätzliche Präzision wurde erreicht zu Lasten einer größeren Anzahl von Annahmen, welche möglicherweise zu zusätzlichen Verzerrungen führen können (siehe hierzu Harrell, 2001).

Es existieren komplexere parametrische Modelle (vgl. Harrell, 2001), welche sich unterschiedlichen Mustern (etwa für frühe und späte Sterblichkeit) anpassen. Solche Modelle besitzen mehr freie Parameter, welche eine größere Adaptation an beobachtete Muster erlauben und sie tendieren dazu, genauere Schätzungen zu produzieren als nicht-parametrische Analysen. Diese komplexen parametrischen Modelle haben den Nachteil, dass die Überlebenskurve an einem bestimmten Punkt möglicherweise substantiell von zeitlich entfernten Ereignissen beeinflusst werden könnte, und besonders mag dies dazu verleiten, die Kurve hinter jene Region zu extrapolieren.

Hazardfunktionen (hazard functions)

Falls eine große Anzahl erfolgreicher Abstinenzler (Datenbeispiel nach Stevens & Hollis, 1989) plötzlich in einem gegebenen Monat rückfällig wird, fällt die Überlebensfunktion in diesem Bereich scharf ab. Bei einem solchen starken Abfall sind also Ex-Raucher mehr gefährdet, einen Rückfall zu erleiden. Es ist demnach eine Möglichkeit, risikoreiche Zeitperioden zu identifizieren, indem Abschnitte mit starken Steigungsänderungen aus der Überlebensfunktion isoliert werden. Ein besserer Weg dies zu tun, ist jedoch, die *Hazardfunktion* zu betrachten. Die Hazardfunktion oder *integrierte Ausfallrate* ist eine verwandte mathematische Funktion, die die Veränderungen der Steigung der (log) Überlebensfunktion registriert.

Mathematische Definitionen von „Hazard“, respektive der „Ausfallrate“ unterscheiden sich abhängig davon, ob die Zeit diskret oder kontinuierlich gemessen wird. Wird die Zeit *diskret* gemessen, ist die Ausfallrate die bedingte Wahrscheinlichkeit, dass ein Ex-Raucher innerhalb eines bestimmten Zeitintervalls rückfällig wird, vorausgesetzt, dass diese Person zu Beginn des entsprechenden Intervalls noch nicht rückfällig war. Mit abnehmender Intervall-Länge nimmt auch die Wahrscheinlichkeit ab, dass das fragliche Ereignis während jeglichen gegebenen Intervalls eintritt.

Wird die Zeit *kontinuierlich* gemessen, muss die Definition der Ausfallrate modifiziert werden, da die Wahrscheinlichkeit, dass ein Ereignis zu irgendeinem Moment eintritt, sich Null nähert. Demnach ist in kontinuierlicher Zeit die Ausfallrate die augenblickliche Rate des Rückfalls, unter der Annahme von ununterbrochener Abstinenz bis zu dieser Zeit.

Ähnlich wie die Überlebensfunktion kann die Hazardfunktion gegen die Zeit abgetragen werden, was zu einem Risikoprofil des Rückfalls für die jeweilige Zeit führt (hier: für jede Woche).

Die Größe der integrierten Ausfallrate repräsentiert das Risiko, in dieser Zeit rückfällig zu werden: Je größer die Ausfallrate, desto größer das Risiko. Die Ausfallrate in jedem Zeitintervall wird berechnet, indem nur Daten derjenigen Personen mit einbezogen werden, die immer noch das Ereignis erleben können (also bis dahin noch nicht rückfällig geworden sind) – die sogenannte *Risikogruppe* (*risk set*).

Die *Abbildung 5* zeigt die zugehörige Hazardfunktion zu der Überlebensfunktion in Abbildung 3. Das Risiko, einen Rückfall zu erleben, ist in jeder der ersten Wochen der Studie hoch und geht dann mit zunehmender Zeit zurück. Ehemalige Raucher haben folglich ein größeres Risiko, unmittelbar nachdem sie das Rauchen aufgegeben haben, rückfällig zu werden. Diejenigen Raucher hingegen, welche erfolgreich seit mehreren Monaten abstinent sind, bleiben wahrscheinlich auch für mindestens ein Jahr „clean“.

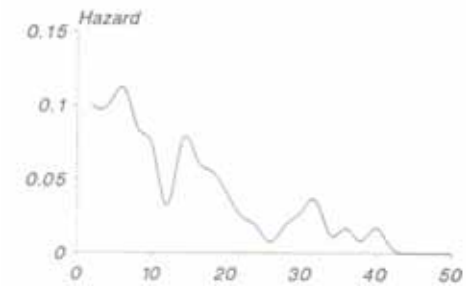


Abbildung 5: Zugehörige Hazardfunktion zu der Überlebensfunktion in Abb. 3 (entnommen aus: Singer & Willett, 1991, p.270).

Ein wichtiger Kritikpunkt an der Überlebensfunktion ist der, dass sie als Zusammenfassung der Daten eine zu starke Vereinfachung darstellt, aufgrund ihrer konsistenten Form, ungeachtet der Verteilung des Risikos. Im Gegensatz hierzu erfasst die Hazardfunktion effektiv die Verteilung des Risikos über die Zeit.

Es lässt sich auch hier mühelos eine Parallele zu einem Beispiel aus dem universitären Kontext ziehen. Scheidet eine große Anzahl von Studierenden plötzlich in einem bestimmten Semester aus dem Studium aus, fällt die Überlebensfunktion in diesem Bereich verstärkt ab. In dieser Zeit sind also Studierende mehr gefährdet, das Studium abzubrechen. Betrachtet man nun die zugehörige Hazardfunktion, so lassen sich risikoreiche Zeitabschnitte leicht (optisch) identifizieren. Es entsteht ein Risikoprofil des Studienabbruchs für die jeweilige Zeit (hier beispielsweise für jedes Semester). Je größer die Ausfallrate, desto größer das Risiko. Die Ausfallrate in jedem Intervall bezieht nur Daten derjenigen Personen ein, die immer noch das Ereignis erleben können (also noch nicht abgebrochen haben) – die Risikogruppe.

Abbildung 6 zeigt vier Hazardfunktionen, von denen jede ein verschiedenes Risikomuster aufzeigt. Da die „Spitzen" (*peaks*) in den Funktionen auf Perioden erhöhten Risikos hinweisen, zeigen sie an, wann das Zielereignis am wahrscheinlichsten auftritt.

Hazardfunktion *A* ist flach; das Risiko ist unabhängig von der Zeit und das Ereignis tritt zufällig auf. Weil unter anderem Alter, Zeit und Kohorte menschliches Verhalten beeinflussen, sind flache Hazardfunktionen eher selten in psychologischen Untersuchungen. Nichtsdestotrotz wurde (zeit-)dauer-unabhängiges Verhalten in Studien gefunden, die die Zeit bis zur Scheidung nach der Geburt eines Kindes (Fergusson, Horwood & Shannon, 1984; nach Singer & Willett, 1991) oder die Zeit bis zu Aufmerksamkeitswechseln im Klassenraum (Felmlee & Eder, 1983; nach Singer & Willett, 1991) untersuchten.

Hazardfunktionen mit einem deutlichen Gipfel sind hingegen häufiger zu finden. Tritt das Zielereignis am wahrscheinlichsten unmittelbar nachdem „die Uhr startet" auf, so hat die Funktion ihren „peak" früh. Solche Funktionen finden sich beispielsweise in Sucht-Rückfallsstudien, aber auch in Studien, die die Wiederkehr von ehelicher Gewalt untersuchen (z.B. Berk & Sherman, 1985, 1988; nach Singer & Willett, 1991).

Andere Hazardfunktionen haben ihren „peak" in der Mitte oder zu späten Zeitpunkten. Steigt das Risiko, eine Begebenheit zu erleben mit der Zeit, wird die Hazardfunktion ihren Gipfel eher später haben. Solche Funktionen wurden in vielen Bereichen gefunden, einschließlich in Studien, die die menschliche Lebenszeit untersuchen (Gross & Clark, 1975; nach Singer & Willett, 1991) und den Eintritt in den Ruhestand (Campbell, Mutran & Parker, 1987; nach Singer & Willett, 1991).

Mittlere Gipfel finden sich überwiegend in Studien mit langen Datenerhebungsperioden. Diese Designabhängigkeit entsteht wegen eines einfachen Grundes: In kurzen Studien erscheint die bestimmte Zeit, in der sich ein Gipfel findet, auf der Zeitachse spät, da die Datenerhebung hier endet. Daraus folgt, dass in retrospektiven Studien, die häufig lange Perioden von Zeit abdecken, oft solche Hazardfunktionen aufgedeckt werden.

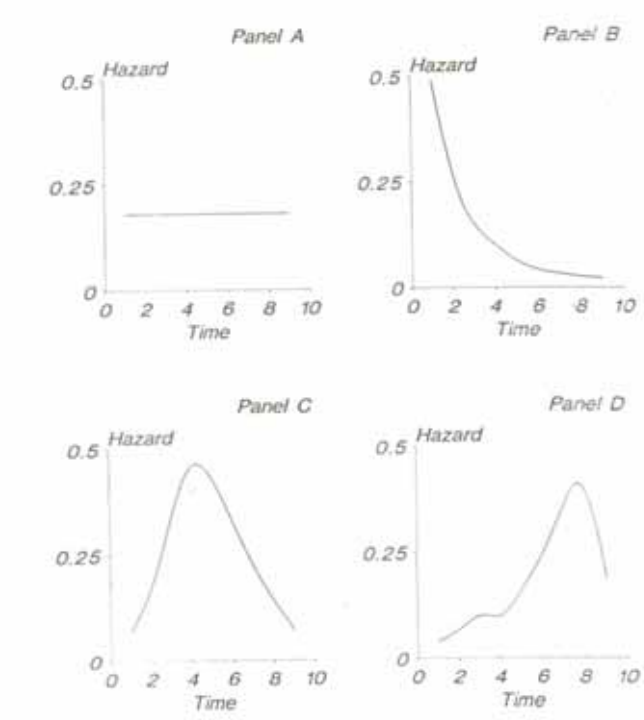


Abbildung 6: Vier prototypische Hazardfunktionen. Panel A: flach; Panel B: früher Gipfel; Panel C: mittlerer Gipfel; Panel D: später Gipfel (entnommen aus: Singer & Willett, 1991, p.271).

Es ist oftmals zweckmäßiger und besser interpretierbar, anstatt mit Überlebenskurven, direkt mit den Ausfallraten (*Hazard*: das Risiko, dass ein Ereignis pro vergangener Zeiteinheit auftritt, vorausgesetzt, dass die Person bis zu diesem Zeitpunkt „überlebt“ hat) zu arbeiten. Die Hazardfunktion drückt die Ausfallrate und ihre Veränderungen über die Zeit aus und enthält exakt dieselben Informationen wie die Überlebensfunktion, allerdings in Begriffen einer *Veränderungsrate*. Wo die Survivalfunktion schnell abfällt, ist die Ausfallrate hoch, während bei einer flachen Überlebenskurve die Ausfallrate Null ist. Diese Äquivalenz erlaubt es, dass die Hazardfunktion durch eine mathematische Transformation der Überlebensfunktion abgeleitet werden kann (siehe oben). Bull & Spiegelhalter (1997; p. 1052) zeigen, dass eine glatte, parametrische Weibull-Überlebenskurve in eine glatte Hazardfunktion transformiert wird (vgl. Abb. 2). Eine identische Transformation kann auf diejenigen Überlebensfunktionen angewandt werden, die mittels der Kaplan-Meier-Methode generiert wurden. Hier erscheint dann allerdings die Hazardfunktion als eine Serie von Spikes, da die K-M-Überlebensfunktion in diskreten Stufen treppenförmig fällt, und sie muss generell geglättet werden, um eine gut interpretierbare Graphik darzustellen (siehe hierzu Bull & Spiegelhalter, 1997).

Eine Analogie

Da Hazard- und Überlebensfunktionen eher unvertraute Konzepte darstellen, lässt sich zur Veranschaulichung eine epidemiologische Analogie zu den Konzepten von Inzidenz und Prävalenz ziehen. Die *Inzidenz(rate)* erfasst die Anzahl der neuen Ereignisse, die während eines Zeitintervalls (ausgedrückt als Verhältnis zu der Zahl der „gefährdeten" Individuen) auftreten, wohingegen die *Prävalenz(rate)* diese Risiken kumuliert zu der totalen Anzahl von Ereignissen, die bis zu einer gegebenen Zeit aufgetreten sind. Inzidenz und Prävalenz korrespondieren direkt mit Hazard und Survival: *Hazard repräsentiert Inzidenz und Survival repräsentiert die kumulierte Prävalenz*.

Diese Analogie macht die Wichtigkeit deutlich, *beide* Funktionen (Hazard- und Überlebensfunktion) zu betrachten. Epidemiologen haben schon lange bemerkt, dass obwohl die Prävalenz(rate) das Ausmaß eines Problems zu einem bestimmten Zeitpunkt schätzen kann, die Inzidenz(rate) der Schlüssel zur Ätiologie einer Erkrankung ist. Prävalenz konfundiert nämlich Inzidenz mit (zeitlicher) Dauer. Zustände mit längerer Dauer mögen eine höhere Prävalenz haben, obgleich sie ähnliche oder niedrigere Inzidenzraten haben. Um zu bestimmen, wann Personen gefährdet sind, betrachten Epidemiologen die Inzidenz (Hazard).

Ausblick

In dieser Einführung wurden die Grundzüge der Survival Analysis dargestellt, und zwar ohne Annahme explanatorischer Variablen. Die Methoden der Survival Analysis sind aber unter anderem auch in der Lage, (eine oder mehrere) explanatorische Variablen mit einzubeziehen. So ist es zum Beispiel denkbar, eine Stichprobe in Subgruppen aufzuteilen und diese miteinander zu vergleichen. Der Untersucher könnte vielleicht das Geschlecht, sozioökonomischen Status, Studienzufriedenheit o.ä. als eine explanatorische Variable im Zusammenhang mit Studienabbruchsraten untersuchen wollen. Natürlich kann man auch hier etwa Kaplan-Meier-Schätzungen für beide Subgruppen erstellen, und miteinander vergleichen. Auch die Ausfallraten für beide Gruppen können (analog den Ausführungen oben) berechnet und verglichen werden. Folgen können dann statistische Signifikanztestungen der Unterschiede zwischen den Kaplan-Meier-Funktionen (für eine Vertiefung siehe Bull & Spiegelhalter, 1997). So kann die Survival Analysis maßgeblich dazu beitragen, prognostische Faktoren zu verstehen (Form und Stärke), und substantielle Effekte von Einflussfaktoren sowie Interaktionen zwischen Variablen zu beleuchten.

Literatur

Allison, P.D. & Liker, J.K. (1982). Analyzing sequential categorial data on dyadic interaction: A Comment on Gottman. *Psychological Bulletin, 91,* 393-403.

Berk, R.A. & Sherman, L.W. (1985). Data collection strategies in the Minneapolis domestic assault experiment. In: L. Burstein, H.E. Freeman & P.H. Rossi (Eds.), C*ollecting evaluation data: Problems and solutions* (pp. 35-48). Beverly Hills, CA: Sage.

Berk, R.A. & Sherman, L.W. (1988). Police responses to family violence incidents: An analysis of an experimental design with incomplete randomisation. *Journal of the American Statistical Association, 83*, 70-76.

Bull, K. & Spiegelhalter, D.J. (1997). Tutorial in Biostatistics: Survival Analysis in Observational Studies. *Statistics in Medicine, 16,* 1041-1074.

Campbell, R.T., Mutran, E. & Parker, R.N. (1987). Longitudinal design and longitudinal analysis: A comparison of three approaches. *Research on Aging, 8*, 480-504.

Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, 34,* 187-202.

Daniel, H.-D. (1996). Korrelate der Fachstudiendauer von Betriebswirten: Ergebnisse einer Absolventenbefragung an der Universität Mannheim. *Zeitschrift für Betriebswirtschaft, 1* (E), 95-115.

Felmlee, D. & Eder, D. (1983). Contextual effects in the classroom: The impact of ability groups on student attention. *Sociology of Education, 56,* 77-87.

Fergusson, D.M., Horwood, L.J. & Shannon, F.T. (1984). A proportional hazards model of family breakdown. *Journal of Marriage and the Family, 46,* 539-549.

Fichmann, M. (1988). Motivational consequences of absence and attendance: Proportional hazard estimation of a dynamic motivation model. *Journal of Applied Psychology, 73*, 119-134.

Giesen, H. & Gold, A. (1996). Individuelle Determinanten der Studiendauer. Ergebnisse einer Längsschnittuntersuchung. In: J. Lompscher & H. Mandl (Hrsg.), Lehr- und *Lernprobleme im Studium* (S. 86-99). Bern: Huber.

Gold, A. & Kloft, C. (1991). Der Studienabbruch: Eine Analyse von Bedingungen und Begründungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 23*, 265-279.

Gross, A.J. & Clark, V.A. (1975). *Survival distributions*. New York: Wiley.

Harrell, Jr., F.E. (2001). *Regression Modeling Strategies*. New York: Springer.

Hunt, W.A., Barnett, W. & Branch, L.G. (1971). Relapse rates in addiction programs. *Journal of Clinical Psychology, 27,* 455-456.

Hunt, W.A. & General, W.R. (1973). Relapse rates after treatment for alcoholism. *Journal of Community Psychology, 1,* 66-68.

Hutchinson, D. (1988). Event history and survival analysis in the social sciences, I: Background and introduction. *Quality and Quantity, 22,* 203-219.

Reissert, R. (1991). Fachstudiendauer: Ist das Problem schon fixiert, welche Handlungsmöglichkeiten gibt es? In: W.-D. Webler & H.-U. Otto (Hrsg.), *Der Ort der Lehre in der Hochschule* (S. 29-60). Weinheim: Deutscher Studienverlag.

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.

Schoenfeld, D.A. & Richter, J.R. (1982). Nomograms for calculating the number of patience needed for a clinical trial with survival as an endpoint. *Biometrics, 38,* 163-170.

Singer, J.D. & Willett, J.B. (1991). Modeling the Days of our Lives: Using Survival Analysis When Designing and Analyzing Longitudinal Studies of Duration and the Timing of Events. *Psychological Bulletin, 110*, 268-290.

Stevens, V.J. & Hollis, J.F. (1989). Preventing smoking relapse using an individually tailored skills training technique. *Journal of Consulting and Clinical Psychology, 57,* 420-424.

Wagemann, C.-H. (1987). Die Schnellen und die Superschnellen. *Hochschulausbildung*, *5*, 115-122.

Klassifikations- und Regressionsbäume

Catherine Myers und Simone Fucks

Einleitung

Welches Verfahren zur Auswahl von Studierenden soll angewendet werden und welche Kriterien gilt es zu beachten, um die besten Studenten und Studentinnen zu selektieren? Vor dieser Frage steht die Mehrzahl der Hochschulen in Deutschland im kommenden Wintersemester 2005/06, wenn die Verantwortung der Bewerberauswahl von der Zentralen Vergabestelle für Studienbewerber (ZVS) zu 60% an die Universitäten weitergeleitet wird. Es gilt zum einen die richtigen Kriterien auszuwählen, die den Studienerfolg und auch den späteren beruflichen Erfolg zuverlässig prognostizieren, zum anderen aber darum, ein entsprechendes Verfahren zu entwickeln, welches diese Kriterien zuverlässig erfasst.

Eine Methode, um eine prognostische Entscheidung zu fällen, stellen Regressions- und Klassifikationsbäume dar. Sie sind in der Lage, Vorhersagen aufgrund von aktuellen Befunden zu machen.

Nach Lefering stellen Klassifikations- und Regressionsbäume (engl. CART, für Classification And Regression Trees) spezielle Verfahren zur Klassifikation von Objekten, in der Regel Personen, dar. Es handelt sich um einen Versuch, intuitiv als sinnvoll erachtete, differenzierte Betrachtungen der Gesamtdaten mit Hilfe eines Algorithmus zu formalisieren (1996).

Obwohl man die Klassifikationsbaum-Analyse noch nicht als ein Standardverfahren bezeichnen kann, wird sie in letzter Zeit zunehmend eingesetzt. Sie findet zum Beispiel Anwendung im medizinischen Bereich, bei der Prognose von Herzinfarkten, bei der Beurteilung der Überlebensrate von Krebserkrankungen und bei der Klassifikation von Alkoholikern (Lefering, 1996).

Aber auch außerhalb der Medizin wird die Klassifikationsanalyse vermehrt angewendet: Sie werden in Assessment-Centern verwendet, um die Auswahl von Bewerbern zu unterstützen. Weitere Bereiche, in denen Baumdiagramme verwendet werden, finden sich in der Werbung und Produktforschung, wenn zum Beispiel festgestellt werden soll, welche Zielgruppen auf bestimmte Produkte ansprechen. In der Wahlforschung wird anhand von Klassifikationsbäume untersucht, welche Merkmale wie Alter, Geschlecht, Beruf und Religionszugehörigkeit sich auf die Wahlentscheidung auswirken. Auch bei der Bonitätsprüfung spielen Klassifikationsanalysen eine wichtige Rolle. Getestet wird auf der Basis von Alter, Einkommen, Besitz und Schulden, ob sich eine Person als kreditwürdig erweist (Bühl, 2002).

Die Vorgehensweise, Daten spezieller Patientengruppen herauszugreifen und für diese Aussagen zu machen, ist weit verbreitet und wurde immer schon praktiziert. Dennoch darf das Risiko der Fehlinterpretation nicht unterschätzt werden. Insbesondere bei kleinen Gruppengrößen und willkürlichen Definitionen der Gruppen ist die Gefahr groß, dass ein beobachteter Zusammenhang nur auf einem Zufall beruht. Um dieser Problematik vorzubeugen, hat man systematische Verfahren zur Subgruppenanalyse eingeführt.

Im Folgenden wird zunächst auf die Konstruktion von Klassifikations- und Regressionsbäume eingegangen, auf Validitäts- und Gütekriterien und auf die Vor- und Nachteile der CART-Verfahren.

Konstruktion von Klassifikationsbäumen

Grundlagen

Die Klassifikationsbaumanalyse ist universell einsetzbar, die Ergebnisse sind leicht zu interpretieren und es werden nur wenige Bedingungen an die Variablen gestellt. Um den unterschiedlichen Anforderungen, die zu analysierenden Datengruppen mit sich bringen, zu entsprechen, bietet das Software Programm SPSS Answertree 3.0 unterschiedliche Algorithmen, die zwischen den Skalenniveaus der Variablen unterscheiden. SPSS Answertree 3.0 unterscheidet zwischen dem CHAID-, exhaustive CHAID-, C&RT- und Quest-Algorithmus. Alle vier stellen keine Anforderungen an die Prädiktorvariable, auch die abhängige Variable kann bei allen nominal, ordinal oder metrisch skaliert sein, lediglich der Quest-Algorithmus wird nur für nominale Zielvariablen eingesetzt (Bühl, 2002).

Zur Konstruktion von Klassifikationsbäumen müssen zuerst die Einflussgrößen (UV) und die Zielgrößen (AV) definiert und operationalisiert werden. Aus einer Untersuchung von Reiß und Moosbrugger beispielsweise geht hervor, dass die Vordiplomsnote, die Abiturnote sowie EDV-Kenntnisse gute Prädiktoren für ein erfolgreiches Studium darstellen. Als abhängige Variable dienen Studienleistungen und Studiendauer, die anhand der Hauptdiplomsnote und der Anzahl der Semester operationalisiert wurden (2002). Die Festlegung von Einflussgrößen beruht häufig auf Erfahrungswerten von Fachleuten und entspricht daher einem heuristischen Verfahren.

Obwohl für Klassifikationsbäume keine grundlegenden Voraussetzungen an die Art der Variable gestellt werden, ist die Form der Variable für die weitere Verfahrensweise relevant. Daher wird zwischen dichotomen, nominalen, kategorialen, ordinalen und stetigen Variablen unterschieden. Von der Gestalt der Daten hängt die Wahl der statistischen Kenngrößen ab, mit denen die Beobachtung zusammenfassend beschrieben wird.

Konstruktion eines Klassifikationsbaumes

Zur Vereinfachung der Darstellung wird sich im Folgenden auf binäre Klassifikationsbäume beschränkt.

Die Konstruktion von binären Klassifikationsbäumen führt zu einer Teilung (Split s) an einem Merkmal und resultiert in der Entstehung von zwei Untergruppen. Es folgt eine fortlaufende Verzweigung in immer kleiner werdende Untergruppen, welche als rekursiv bezeichnet wird. Analog lässt sich dies mit dem Bild eines umgekehrten Baumes vergleichen. Die Gesamtpopulation der Probanden stellt den Stamm dar, der sich an einem bestimmten Merkmal in zwei Hauptäste verzweigt und sich in immer kleinere Zweige verästelt. Eine Gruppe, die nicht weiter aufgeteilt wird, bezeichnet man analog zum Astende eines Baumes als Blatt. Es handelt sich um eine nicht-endende Subgruppe, die sich noch weiter teilt und die Quadrate zeigen an, dass es sich um endende, terminale Subgruppen handelt, die als Endknoten oder bildlich als Blätter bezeichnet werden. Die Splits erfolgen, nachdem für die jeweilige Teilung s Konditionen definiert werden. So könnte beispielsweise am ersten Split erfragt werden, ob die Studienbewerberin oder der Studienbewerber einen Abiturschnitt unter oder über 2,0 hat.

$x_2 = \{x; x_1 < 2{,}0\}$, $x_3 = \{x; x_1 > 2{,}0\}$

Falls die Person darunter liegt, wird sie in x_2 eingruppiert, im anderen Fall wird sie x_3 zugeteilt. Sobald die Testperson eine Endgruppe erreicht hat, wird ihr eine entsprechenden Klassifikationsgruppe zugeteilt. Eine genauere Erklärung hierzu findet sich im Abschnitt zu den Klassifikationsregeln.

Das Ergebnis einer CART- Analyse ermöglicht eine sehr einfache Interpretation und Umsetzung für die Prognose zukünftiger Fälle: Eine Bewerberin oder ein Bewerber durchläuft den Baum, indem er den Verzweigungen entsprechend seinen individuellen Werten der Einflussgrößen folgt, bis er schließlich in einem der Blätter endet. Die Schwierigkeit bei der Erstellung von Klassifikationsbäumen besteht darin, geeignete Splits festzulegen, zu bestimmen, wann die Teilung gestoppt werden muss und in welche Klassifikation eine Endgruppe einzuteilen ist.

Split-Regeln

Anzahl an Splits

Eine der großen Herausforderungen bei der Konstruktion von Klassifikationsbäumen liegt darin, die richtige Größe des Baumes zu finden, um optimale Ergebnisse zu produzieren.

Generell führt eine größere Anzahl von Teilungen zu einer Reduktion der Missklassifikationen und somit zu einer höheren Güte des Modells. Wird sooft geteilt, dass jeder Endknoten nur noch genau einen Fall enthält, so beträgt die Missklassifikationsrate 0.

Außerdem wird bei zu kleinen Bäumen die komplette Information nicht genügend ausgeschöpft.

Auf der anderen Seite führen zu große Bäume ab einer gewissen Größe zu einem Informationsüberdruss und können sogar zu einer Erhöhung der Fehlklassifikation führen. Die Missklassifikationsrate nimmt nämlich sowohl bei zu kleinen als auch bei zu großen Bäumen zu. Des Weiteren werden zu große Bäume unübersichtlich und resultieren in Fehlinterpretationen.

Es empfiehlt sich daher einen moderaten Baum zu erstellen, zum einen aufgrund von Erfahrungswerten aus vorangegangen Modellen, zum anderen Anhand der Ergebnissen der Missklassifikationsraten. Ein weiteres Hilfsmittel stellt das sogenannte Pruningverfahren dar, welches im gleichnamigen Abschnitt näher erläutert wird.

Bester Split

Nach der Bestimmung der Anzahl der Splits, stellt sich im Anschluss daran das Problem, einen optimalen Split zu ermitteln. Eine Teilung sollte derart erfolgen, dass die daraus entstehenden Untergruppen „reinere“ Daten liefern als die vorherige Subgruppe oder umgekehrt, sollte bei jeder Trennung der Grad der „Unreinheit“ abnehmen. So möge in dem Beispiel mit den Studienbewerbern nach der ersten Teilung bezüglich der Abitursnoten, die daraus resultierenden Subgruppen mehr Informationen über die Eignung als Psychologiestudent(in) enthalten als vorher. Das heißt also, dass die Personen mit einer Abitursnote unter 2,0 geeigneter für das Studium sein sollten als die Bewerber(innen) mit einer schlechteren Abschlussnote. Dieser Informationszuwachs lässt sich dadurch feststellen, dass sich die beiden durch die Teilung entstanden Subgruppen möglichst stark von dem übergeordneten Vaterknoten unterscheiden sollten.

Die Unterscheidung der Subgruppe von der vorangehenden Gruppe kann auf zweierlei Art bestimmt werden. Zum einen erfolgt dies durch das Chi-Quadrat-Kriterium, zum anderen mit der Bewertungsfunktion i (node impurity function).

Zunächst sei das Prinzip des Chi-Quadrat Kriteriums näher erläutert. Mit n Personen und einer Prävalenz der Zielgruppe von p ($0<p<1$) ergibt ein Split eine Vierfeldertafel, mit den

Randsummen definiert durch die Prävalenz einerseits und durch die Variable andererseits, die den Split induziert. Die Auftrennung hat dann für die Zielgruppe keine Bedeutung, wenn man bei der linken und rechten Subgruppe etwa dieselbe Prävalenz p im Ausgangsknoten beobachtet. „Je stärker die beobachteten Prävalenzen voneinander abweichen, desto wichtiger ist dieser Split für die Prognose. Der Chi-Quadrat misst diese Abweichung, indem er die Quadrate der Abweichungen relativ zur Zahl der erwarteten Ereignisse aufsummiert (Lefering, 1996)“.

$$\text{Chi}^2 = \sum \frac{(\text{beobachtet} - \text{erwartet})2}{\text{erwartet}} \qquad \text{(Gl. 1)}$$

Ein weiteres Verfahren zur Festlegung des besten Splits erfolgt anhand einer allgemein definierten Bewertungsfunktion i (node impurity functions), welche auf Breimann et al. und Gordon et al. zurückgeht. Da das Ziel eines jeden Splits auf eine Erhöhung des Reinheitsgrades abzielt, wurde dazu eine Bewertungsfunktion i hergeleitet, die von 1 schlecht bis 0 optimal reicht.

Jeder Split führt dabei zu einer Verminderung der Unreinheit. Der beste Split wird dann anhand des Grades der Verbesserung durch die Funktion i bestimmt. Das Maß ist die Differenz der Bewertung des Baumes mit und ohne diesen Split. Rechnerisch ergibt sich somit folgende Formel:

$$\Delta i\,(s, t) = i\,(t) - p_R i(t_R) - p_L i(t_L) \qquad \text{(Gl. 2)}$$

Aus Gleichung 2 läßt sich ersehen, dass an einem Split s ein Teil p_R eines Knoten in die rechte Baumhälfte t_R und der andere Teil p_L in die linke Hälfte t_L geordnet wird.

Es zeigt sich, dass je weiter die beiden Subgruppen voneinander abweichen, desto geringer wird die „Unreinheit “ i (impurity) und der Split umso besser.

Beide Verfahrensweisen haben inhaltlich ihre Berechtigung. Das Chi-Quadrat-Kriterium ermittelt unabhängig von vorangehenden Splits die optimale Teilung, während die Bewertungsfunktion i von der globalen Prävalenz eines Splits abhängt. Das Ergebnis des einen Verfahrens ist daher nicht notwendig mit dem des anderen Verfahrens identisch.

Die Daten sind am reinste, wenn jede Endgruppe zu nur einer Klassifikation führt. Der geringste Informationsgewinn liegt dann vor und damit auch die „unreinsten“ Daten, wenn in jeder Endgruppe, sowohl geeignete als auch ungeeignete Bewerber(innen) auftreten können.

Unbedingt gilt es bei der Wahl des besten Splits die Anzahl der Missklassifikationen zu berücksichtigen und die Kosten der Missklassifikationen. Die Missklassifikationsrate R (d) wird am häufigsten dadurch ermittelt, indem eine Lernstichprobe L anhand eines Klassifikationsbaumes analysiert wird. In einem zweiten Durchgang wird das ermittelte

Konstrukt d dann nochmals an der gleichen Stichprobe validiert. Dieser Vorgang wird auch als resubstitution estimate bezeichnet. Die dazu gehörige Klassifikationsrate R(d) wird durch folgende Gleichung bestimmt:

$$R(d) = 1/N \Sigma (X(d(x) \neq j) \quad \text{(Gl. 3)}$$

Es wird in Gleichung 3 bestimmt, wie oft eine Person fälschlicherweise in Klasse j eingeordnet wird. Die Gleichung kann somit entweder 1 werden, falls die Person falsch klassifiziert wurde, andernfalls 0. Dieses Verfahren tendiert jedoch dazu, die Missklassifikationsrate zu unterschätzen, da sie an der gleichen Stichprobe validiert wird, anstelle einer unabhängigen Stichprobe. Trotzdem wird dies aufgrund von fehlenden Stichprobengrößen am häufigsten eingesetzt. Alternative Verfahren werden im Abschnitt über Validität behandelt. Die Kosten einer Fehlklassifikation werden definiert als die Kosten, die entstehen, wenn ein Objekt, welches eigentlich zur Klasse j gehört, fälschlicherweise in die Klasse i eingeordnet wird.

Wird ein Objekt, welches im Endknoten t endet, einer Klasse i zugeordnet, so lauten die Kosten einer Fehlklassifikation:

$$\sum c(i/j) \cdot p(j/t). \quad \text{(Gl. 4)}$$

Die Kosten von Missklassifikationen stellen einen sehr wichtigen Aspekt dar, den es bei der Bestimmung eines optimalen Splits unbedingt zu berücksichtigen gilt. Um die Tragweite bei Missachtung der Kosten zu verdeutlichen, sei dies am Beispiel aus der Medizin verdeutlicht. Betrage die Mortalitätsrate nach einem Herzinfarkt bei einem Patienten 20%, und die Wahrscheinlichkeit zu überleben 80%, so ist es unbedingt zu beachten, die Patienten mit dem höchsten Sterberisiko herauszufiltern, denn es sind diejenigen, die höchste Pflege und aufwendigste Therapien benötigen. Eine Nichtbeachtung der Kosten der Missklassifikation könnte dann in diesem Beispiel fatale Folgen haben. Herabsenken lässt sich dieses Risiko durch eine Erhöhung von Teilung und durch das Festlegen von Priors, auf die in einem späteren Abschnitt näher eingegangen wird.

Das zwei Klassen Problem

In den vorangehenden Abschnitten wurde der optimale Split anhand des Chi-Quadrat Kriteriums und der Bewertungsfunktion i gemessen. Eine alleinige Anlehnung an die beiden Verfahren zur Ermittlung des besten Splits genügt jedoch nicht. Dies soll im folgenden verdeutlicht werden. Setzen wir die Bewertungsfunktion i durch die Anzahl der Missklassifikationen R (d) fest, d, h, der beste Split entspricht der Teilung, welche die Missklassifikationsrate am meisten reduziert, lassen sich zwei Defizite erkennen.

Der erste Mangel besteht darin, dass sowohl die vorangegangene Gruppe Klasse 1 favorisiert, als auch die Mehrheit der beiden daraus resultierenden Subgruppen Klasse 1 bevorzugen, beziehungsweise eine Gleichheit besteht.. Abbildung 1 verdeutlicht diese Problematik.

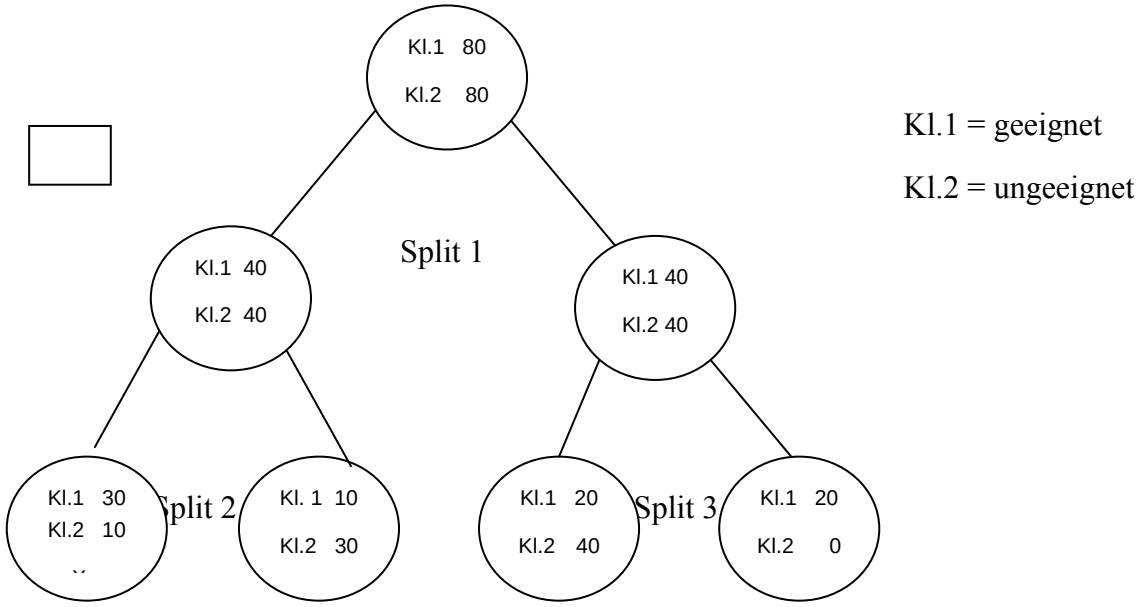


Abbildung 1

Darstellung eines Baumdiagramms zur Verdeutlichung des Zwei-Klassen Problems

Klasse 1 steht beispielsweise für die Anzahl an geeigneten Bewerber für das Studienfach Psychologie, während Klasse 2 für nicht geeignet befunden wurde. Die Graphik zeigt deutlich, dass durch die erste Teilung kein Zugewinn an Informationen erreicht wird. Nach Gleichung 2, beträgt daher der Zuwachs an Reinheit 0, weil die 800 geeigneten Bewerber (Klasse 1) symmetrisch in die beiden Subgruppen aufgeteilt werden. Diese unterscheiden sich somit nicht von dem vorangegangenen Knoten. Analog betrifft dies auch die ungeeigneten Probanden- innen in Knoten x_3. Auch die Berechnung der Missklassifikationsrate nach Gleichung 2, welche ein Ergebnis von 1 erbringt, bestätigt, dass es sich hierbei um keinen optimalen Split handelt.

Das zweite Problem ist etwas schwieriger zu quantifizieren. Aus Abbildung 1 läßt sich ebenfalls entnehmen, dass sich sowohl bei Split 2 als auch bei Split 3 eine Missklassifikationsrate von R (T) = 0.25 ergab, weil jeweils 200 Probanden aus 800 falsch klassifiziert worden sind. Es lässt sich daher nicht vorhersagen, welcher Split zu bevorzugen ist. Außerdem ist aus der Abbildung 1 zu ersehen, dass nach der Formel r(t) = p(j/t), wobei j die Anzahl der Fehlklassifikationen und t die Anzahl der Gesamtfälle im

Knoten betrifft, beim dritten Split der Knoten x_6 eine Missklassifikation von r (t) = 0 auftritt und somit keine weitere Teilung mehr bedarf. Der andere Knoten x_5 mit einer Missklassifikationsrate von r(t) = 0.5 liegt damit wesentlich höher als alle anderen Endknoten, die eine Missklassifikationsrate von r(t) = 0.33 aufweisen. Da Split 3 somit einen perfekten und einen sehr unreinen Knoten enthält, ist er dem zweiten Split nicht vorzuziehen. Die Bewertungsfunktion i führt daher nicht zu einer ausreichenden Anerkennung von reinen Knoten. Diese Problematik nimmt mit zunehmender Größe des Baumes weiter zu. Das zwei-Klassen Problem schafft diesem Problem Abhilfe. Um den besten Split auszuwählen sei daher in diesem Fall folgende Bewertungsfunktion gültig:

$$i(t) = \Phi(p_1) \qquad \text{(Gl. 5)}$$

Es wird von der Annahme ausgegangen, dass beide Wahrscheinlichkeitsvorhersagen symmetrisch verteilt sind. Beträgt $p_1 > 1/2$, nehmen beide Seiten nur linear ab, so dass kein Split zu bevorzugen ist. Um nun einen Baum zu konstruieren, der eindeutig die 3. Teilung in Abbildung 1 bevorzugt, muss $\Phi(p_1)$ hinsichtlich der Missklassifikationsrate schneller abnehmen. Dies wird dadurch erreich, dass Φ konkav verläuft. Es wird somit nicht mehr nach einem Split geschaut, der nur die Anzahl der Fehlklassifikationen reduziert, sondern nach einem Split, der diese am schnellsten vermindert.

Multiklassen Probleme

Multiklassen Probleme sind Generalisierungen der zwei Klassen Probleme und werden als Gini Kriterium und Twoing Kriterium bezeichnet.

Das Gini Kriterium teilt den Endknoten t zufällig einer Klasse mit der Wahrscheinlichkeit p (j/t) zu. Die Wahrscheinlichkeit, dass t zur Klasse i gehört beträgt somit p (i/t). Die daraus resultierende Definition ergibt daher:

$$i(t) = \sum p(j/t)\, p(i/t) \qquad \text{(Gl. 6)}$$

Diese Formel kann auch derart verwendet werden, dass einer Klasse j der Wert 0 zugesprochen wird und der Klasse i der Wert 1. Wird dies in einem Knoten mit jedem Probanden durchgeführt, so ergibt sich eine Stichprobenvarianz mit folgender Form:

$$\sum p(j/t)\,(1- p(j/t)) = 1 - \sum p^2(j/t) \qquad \text{(Gl. 7)}$$

Es gilt hierbei festzustellen, dass es sich um ein Polynom zweiter Ordnung handelt, das eine konkave Form darstellt.

Beim Twoing Problem werden an jedem Knoten die Klassen in zwei Superklassen eingeteilt nach folgender Regel:

$C_1 = \{j_1,.....,j_n\}$, $C_2 = C - C$ (Gl. 8)

Alle Personen, die in C1 eingeteilt werden sind Klasse 1 Probanden, die restlichen gehören zu Klasse 2. Analog dem zwei-Klassen Problem gilt es dann den besten Split zu finden, bei dem die Missklassifikationsrate am schnellsten vermindert wird. Dieses Verfahren hat den Vorteil, dass ein „strategischer" Split gebildet wird, der den Analytiker über ähnliche Klassen informiert. Mit strategisch ist gemeint, dass Klassen zusammengruppiert werden, die in einer bestimmten Weise zusammengehören. An jedem Knoten werden nämlich immer diejenigen Knoten getrennt, die sich am unähnlichsten sind, so dass man einen guten Überblick über sehr ähnliche und weniger ähnliche Klassen erhält.

Der beste Split beim Twoing Problem wird durch den Split erreicht, der Φ (s, t) maximiert, wenn $C_1 = \{j: p(j/t_L) > p(j/t_R)$.

Das Twoing Problem ist das einzige Verfahren, welches nicht mit einer Bewertungsfunktion

i(t) arbeitet. Dies ist allerdings nicht von Nachteil, da es beim Splitting darum geht einen optimalen Baum zu erhalten, was auch bei diesem Verfahren gewährleistet ist.

Sowohl das Twoing- als auch das Gini-Kriterium haben ihre Vorteile. Beim Gini-Kriterium ist darauf zu achten, dass die Anzahl an Variablen und daraus entstehenden Knoten nicht zu groß angelegt wird, da es sonst sehr leicht zu sehr aufwendigen Bäumen kommen kann.

Welches der beiden Kriterien zum Einsatz kommt, hängt immer sehr stark von dem zu analysierenden Problem ab.

Vergangene Untersuchungen zeigten jedoch, dass beide Verfahren zu sehr ähnlichen Bäumen führen und in ihrer Akkuratheit ebenfalls vergleichbar sind. Der einzige Unterschied besteht darin, dass Gini Teilungen bevorzugt, die zu einem kleinen, reinen Knoten und zu einem großen weniger reinen Knoten tendieren, während das Twoing-Kriterium zu gleich großen Knoten führt. Daher ist Gini in dieser Hinsicht überlegen (Breimann, et al, 1993).

Priors

Die Parameter, die bei der Konstruktion eines Baumes festgelegt werden können, sind die Wahrscheinlichkeitsvorhersagen (Priors) und die Kosten der Missklassifikation. Beide Parameter sind nicht unabhängig voneinander.

Wahrscheinlichkeitsaussagen sind sinnvoll bei der Konstruktion eines Baumes, da sie eine genaue Vorhersage und damit ein akkurates Ergebnis unterstützen, denn die Daten in einer Population sind nicht immer symmetrisch verteilt. Gäbe es beispielsweise zehnmal so viele Bewerber (innen), die für ein Psychologiestudium geeignet sind als ungeeignete

Probandinnen (10:1), es würde aber von einer proportionalen Verteilung ausgegangen, läge die Missklassifikationsrate bereits bei 10%. Kalkuliert man solche ungleichen Verhältnisse gleich vorab mit ein, kann man die Missklassifikationsrate auf 5% reduzieren.

Dies resultiert natürlich darin, dass, wenn beide Gruppen gleich oft falsch klassifiziert werden, dies bei den geeigneten Psychologiebewerben zu niedrigeren Missklassifikationsraten führt als bei den ungeeigneten, bei denen die Missklassifikationsrate überproportional gesteigert wird. Dies verdeutlicht, dass der Einsatz von Wahrscheinlichkeitsvorhersagen die Missklassifikationensrate in jede beliebig Richtung beeinflussen kann. Es kommt besonders zum Tragen, wenn die Kosten der Missklassifikationsrate bei der einen Gruppe höher sind als bei der anderen. Priors können somit die Kosten von falschen Klassifikationsraten regulieren.

Sollten die Resultate nach einer Wahl von Wahrscheinlichkeitsvorhersagen zweifelhaft sein, so werden unterschiedliche Ausgangsbäume mit verschiedenen Vorhersagen zunächst ausgetestet.

Stop Regeln

Ein Knoten lässt sich theoretisch beliebig oft aufteilen, bis sich in jeder Endgruppe nur eine Person befindet. Ein solches Verfahren ist jedoch wenig geeignet, da es zum einen zu sehr großen Bäumen führt und zum anderen die Interpretierbarkeit erschwert. Da es sich meistens um probabilistische und nur selten um deterministische Prognosen handelt, schwindet mit jeder Trennung die Anzahl der Beobachtungen und die Sicherheit der probabilistischen Aussage nimmt ab. Je weiter man sich also im Klassifikationsbaum vom Stamm entfernt, desto geringer werden die Fallzahlen und um so unsicherer die Prognosen. Bei einem zu kleinen Baum kann auf der anderen Seite die in den Beobachtungen steckenden prognostischen Untersuchungen nur unzureichend genutzt werden. Daher ist die Festlegung wann eine Gruppe nicht mehr in kleinere Untergruppen aufgeteilt werden sollte, eine sogenannte Stop-Regelung von großer Wichtigkeit. Unter Stop-Regel versteht man, ein Kriterium zu finden, welches einen Knoten als Endknoten deklariert.

Das Ende eines Baumes ist – nach heuristischen Regelungen - dann erreicht, wenn durch die Teilung der Subgruppe kein signifikanter Zuwachs an Informationsgewinn verzeichnet wurde, bzw. wenn sich die Bewertungsfunktion nicht deutlich verbesserte.

Nach mathematischen Regelungen müssen sich die beiden durch einen Split induzierten Subgruppen deutlich von der vorangegangen Gruppe hinsichtlich ihrer Prävalenzen unterscheiden, da die Teilung ansonsten keinen Zugewinn an Informationen erbringt, und die Fallzahlen in den aufgetrennten Subgruppen müssen hinreichend groß sein, um den beobachteten Unterschied der Prävalenzen als einigermaßen vertrauenswürdig zu betrachten. Da das Chi-Quadrat-Kriterium beiden Anforderungen gerecht wird, wird dieses Verfahren am häufigsten für die Festlegung der Stop-Regel eingesetzt. Darüber hinaus

empfiehlt es sich jedoch auch, eine Mindestgruppengröße für den letzten Knoten festzulegen. Außerdem kann eine Splitgruppengröße definiert werden, ab der kein weiterer Split mehr durchgeführt werden sollte.

Breimann et al. schlugen alternativ anstelle der Einführung eines Stops das „Pruning“ vor. Ein zusätzlich von den Autoren verwendetes Verfahren schlägt vor, eine akkurate Schätzung der Missklassifikation R (T) bei den zusammengeführten Endungen vorzunehmen

Pruning

Die Methode des Prunings resultiert in einer reduzierten Anzahl an Untergruppen.

Im ersten Schritt wird ein Baum erstellt, der so groß ist, dass die Endknoten entweder so klein sind, dass eine weitere Teilung unvorstellbar ist oder aber keinen Zugewinn an Reinheit mehr bietet. Der sicherste Weg wäre die Stichprobe solange zu teilen, bis sich in jedem Blatt nur noch ein einziger Fall befindet.

Alle Verzweigungen eines Elternknotens werden auch als Nachkommen bezeichnet, während die vorangehenden Knoten Vorfahren genannt werden. Ein Zweig besteht somit aus einem Elternknoten und all seinen Nachkommen.

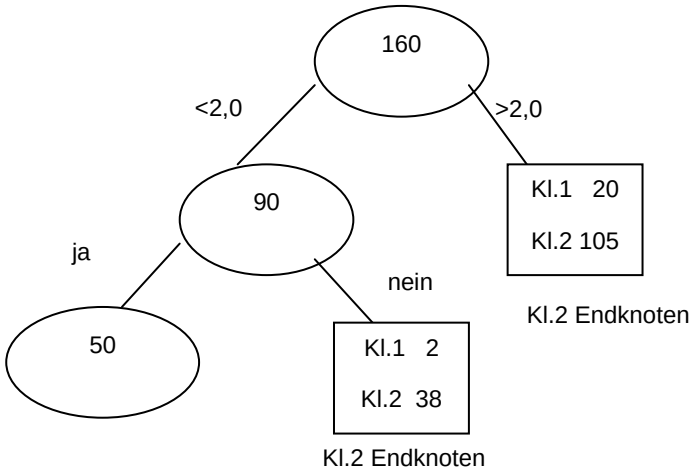
Der Pruning Prozess beginnt ausgehend von dem ursprünglich sehr großen Baum und berechnet für jeden Knoten die Missklassifikationsrate. Im folgenden Schritt werden die Blätter wieder zu ihrem Elternknoten zusammengesetzt, aber immer so, dass sich die Missklassifikation nicht erhöht. Ein solches Pruning wird auch durch die folgende Notation $T\text{-}T_t$ ausgedrückt.

Klassifikationsregeln

Die Klassifikationsregel beschäftigt sich mit der Zuteilung eines Blattes zu einer Klassifikation. Es werden verschiedene Methoden verwendet, um für einen Endknoten die geeignetste Klasse zu finden.

Die von Breimann et al. postulierte „Class Assignment Rule“ ordnet jedem Endknoten t eine Klassifikationsklasse j zu, und zwar derart, dass für jeden Endknoten t diejenige Klasse zugeteilt wird, für welche die Wahrscheinlichkeit p (j/t) am größten ist. Dies entspricht der sogenannten Mehrheitsregel, die besagt, dass t in die Klasse eingestuft wird, in der die meisten Fälle dieses Endknotens verzeichnet wurden. Jeder Person in einem Blatt wird dasjenige Zielereignis zugeordnet, welches die Mehrheit der Personen in diesem Blatt besitzt.

Abbildung 2 verdeutlicht den Sachverhalt.



Kl.1 = geeignet

Kl.2 = ungeeignet

Abbildung 2

Graphische Veranschaulichung der Class Assignment Rule

Wie aus der Graphik zu ersehen ist, wird ein Enknoten also immer derjenigen Klasse zugeteilt, in der die Anzahl des gewünschten Merkmals überwiegt.

Nach der Class Assignment Regel sollte mit der Mehrheitsregel idealerweise auch eine Reduktion der Missklassifikationen einhergehen. Würde dieselbe Person nun wiederholt dem gleichen Baum unterzogen werden, so sollte sie bestmöglichst wieder in die gleiche Endknoten und somit auch in dieselbe Klasse eingestuft werden.

Regressionsbäume

Regressionsbäume sind in ihrer Konstruktion einfacher zu handhaben als Klassifikationsbäume, ansonsten sind sie den Klassifikationsbäumen sehr ähnlich. Sie stellen Konkurrenten der linearen Regression dar und zielen genauso wie die lineare Regression auf eine möglichst minimale Fehlervarianz nach dem Prinzip der kleinsten Quadrate ab.

Der Unterschied zwischen Klassifikations- und Regressionsbäumen besteht darin, dass von vornherein aufgrund von Annahmen oder anhand einer anderen Lernstichprobe eine oder auch mehrere Regressionsgleichungen angenommen werden. Dies legt a-priori y-Werte in jedem Endknoten fest, die innerhalb des Knotens konstant bleiben. Diese y-Werte entsprechen Durchschnittswerten, welche nach dem Prinzip der kleinsten Quadrate die Fehlervarianz minimieren sollen.

Grundlagen

In der linearen Regression werden mehrer Prädiktoren bestimmt, um eine Vorhersage über eine abhängige Variable machen zu können. Die Gleichung hierfür laute wie folgt:

$$y(x) = b_0 + b_1x_1 + b_2x_2 \ldots . + b_mx_m \qquad \text{(Gl. 9)}$$

Der Parameter b muss geschätzt werden.

Ziel ist es den anhand einer Lernstichprobe ermittelten Wert y so zu schätzen, dass er die Werte in einer anderen Stichprobe möglichst genau vorhersagt, und zwar derart, dass die durchschnittliche Abweichung der Fehlerquadratsumme minimiert wird. Analog zu den Klassifikationsbäumen wird auch hier R (d) als ein Gütemaß herangezogen, dass jedoch in diesem Falle folgendermaßen definiert ist:

$$R(d) = \sum(Y - d(x))^2 . \qquad \text{(Gl. 10)}$$

Je kleiner R (d) wird, d.h je genauer die Vorhersage, desto besser ist das aufgestellte Modell.

Bei der Konstruktion von Regressionsbäumen geht es darum eine Form der Variablenauswahl zu treffen. In der linearen Regression entspricht das einer schrittweisen Regression, in der einzelne Variablen extrahiert werden und das Dekrement bestimmt wird.

Interpretation

Obwohl Klassifikationsbäume eine größere Einsicht in die Struktur und Analyse von Daten ermöglichen als die meisten herkömmlichen Verfahren, ist dennoch eine große Sensibilität bei der Exploration, Vorsicht bei der Interpretation und Bedachtheit bei der Entscheidungsfällung geboten.

Die Analyse von Daten bei Entscheidungsbäumen ist eine Mischung zwischen Kunst und Wissenschaft.

Zwei erfahrene Analytiker, denen dieselben Daten präsentiert werden, können zu unterschiedlichen Ergebnissen kommen.

Hierfür gibt es drei unterschiedliche Gründe.

Instabilität der Klassifikationsbäume

Ein Grund für das erwartete Feingefühl bei der Interpretation ist die Instabilität in der Struktur des Klassifikationsbaumes. Viele Variablen enthalten redundante Informationen. Darüber hinaus führen mehrere Splits oftmals zu derselben Verringerung an Reinheit, so dass die Entscheidung zwischen den Teilungen fast zufällig erfolgt. Außerdem kann es vorkommen, dass unterschiedliche Klassifikationsregeln alle annähernd die gleiche Genauigkeit aufweisen, so dass jede von ihnen ausgewählt werden könnte. Dies führt dazu, dass die Bäume häufig sehr unterschiedlich aussehen können.

Nutzung der Ergebnisse

Aufgrund der einfachen Interpretierbarkeit der Ergebnisse, fällt man sehr schnell und unkompliziert eine Entscheidung, ohne zu bemerken, dass der Datensatz die Problematik nur sehr bruchstückhaft repräsentiert.

Um diesem Tatbestand vorzubeugen, sollte immer eine Kreuzvalidierung vorgenommen werden. Darüber hinaus gibt der Grad der Reinheit Aufschluss über die Güte eines Splits und zeigt an, in welche Richtung diese gehen sollen.

Wachstum exploratorischer Bäume

Wie bereits erwähnt ist die Größe eines Baumes von höchster Wichtigkeit, um zu einer richtigen Interpretation zu gelangen. Oftmals ist das Aufstellen eines sehr großen Baumes, um ihn dann zu prunen sehr kostspielig. Außerdem gilt es sich vor der Konstruktion eines Entscheidungsbaumes zu überlegen, ob eventuell einige Merkmale kombiniert werden und welche Wahrscheinlichkeitsvorhersage (Prior) gewählt werden sollen. Ein sehr kostengünstiger und effizienter Weg stellt deshalb die Festsetzung eines Wertes für den Komplexitätsparameter α dar. Das Programm produziert einen optimalen Baum, der mindesten die Anzahl T (α) enthält und nur etwas größer ist.

Ein zweiter Weg ist kleine Untergruppen zu bilden, die es ermöglichen die genau Anzahl an Teilungen in jedem Knoten festzulegen, und es erlaubt nach jeder Teilung die Stop-Regel zu überprüfen.

Konstruktion von Regressionsbäumen

Wie bei den Klassifikationsbäumen erfolgt auch hier anhand eines Merkmals eine Aufteilung in zwei Subgruppen, die sich ihrerseits weiter verästeln, bis sie in einem Blatt enden. In jedem Blatt ist jedoch der y-Wert konstant. Analog zu den Klassifikationsbäumen bestehen drei Aspekte, die für die Konstruktion eines Regressionsbaumes von Wichtigkeit sind.

Im ersten Schritt einen geeigneten Split zu bestimmen, die optimalen Endknoten festzulegen und als letztes, jedem Endknoten einen entsprechenden y-Wert zuzuordnen.

y-Werte

Die y-Werte werden so bestimmt, dass sie die Fehlklassifikationen R (d) möglichst gering halten. Es wird postuliert, dass der Wert, der die Fehlklassifikationen am meisten reduziert dem Durchschnitt der y Werte aller Fälle entspricht.

Der beste Split

Der beste Split bei den Regressionsbäumen ist der Split, in welchem der Fehlerwert R (d) am meisten sinkt. Dies lässt sich durch folgende Gleichung verdeutlichen:

$$R(s, t) = R(t) - R(t_L) - R(t_R). \qquad \text{(Gl. 11)}$$

Dies bedeutet, dass in einem iterativen Verfahren die Knoten so gesplittet werden, dass die Verminderung von R(d) maximiert wird. Dies geschieht, in dem die x-Variable gesucht wird, welche den Vaterknoten am erfolgreichsten in eine Subgruppe mit kleineren Werten und in eine Subgruppe mit größeren Werten aufteilt. Daraus resultiert, dass die durchschnittlichen Werte der Subgruppe $\hat{y}$ (l) kleiner und die durchschnittlichen Werte der Subgruppe $\hat{y}$ (r) größer sind. Anders als bei den Klassifikationsbäumen führt ein Split, der nach Gleichung 11 ermittelt wurde auch tatsächlich zu einer bestmöglichen Teilung.

Stop-Regeln

Wie bei der Klassifikationsanalyse ergibt sich das Problem, die optimale Baumgröße zu bestimmen. Zu große Bäume führen zu sehr unübersichtlichen Graphiken, die zudem sehr schwer zu interpretieren sind, zu kleine Bäume andererseits liefern wenig Informationen. Breimann et al. schlagen daher auch bei den Regressionsbäumen das Pruning vor.

Die einzelnen Knoten werden daher so lange geteilt, bis R (d) minimiert ist. Diese Ausgangsbäume sind in der Regel wesentlich größer als die der Klassifikationsanalyse. Ein Baum gilt in diesem Fall als „rein“, sobald alle y-Werte den gleichen Wert einnehmen. Wenn dies der Fall ist, so werden analog zur Klassifikationsanalyse die einzelnen Äste gekürzt, und zwar derart, dass R (d) nicht größer wird. Das Resultat ist auch in diesem Fall eine optimale Baumgröße.

Klassifikationsregeln

Es wird für den Probanden die Klasse ausgewählt, in welcher der von der Person erreichte y-Werte am wenigsten vom durchschnittlichen y-Wert abweicht. Sollte ein Proband signifikant von dem Durchschnittswert abweichen, so wird er einem anderen Knoten zugeteilt. Ziel ist es, die Varianz in jedem Knoten so gering wie möglich zu halten.

Regressionsbäume werden in der Kriminologie, in der Molekularforschung und im Umweltschutz eingesetzt und sind durchaus konkurrenzfähig mit der linearen Regression.

Ziel sollte es immer sein, einen Tatbestand aus mehreren Blickrichtungen zu betrachten. Regressionsbäume stellen daher eine sehr interessante Alternative dar.

Gütekriterien

Die Güte der Vorhersage wird in der Regel mit der Richtigkeit einer Vorhersage gleichgesetzt. Am Ende der Entwicklung eines solchen Klassifikationsbaumes steht ein Verfahren, das eine Vorhersage machen soll, zum Beispiel, ob eine Bewerberin für das Studienfach Psychologie ausgewählt wurde oder nicht.. Das Modell muss sich also anhand der gegebenen Daten für eines von zwei möglichen Ergebnissen entscheiden. Bei der Erstellung des Klassifikationsbaumes ist es zwingend notwendig, die Güte des Baumes in Hinblick auf die Prognostizierbarkeit der Ergebnisse zu beurteilen. Da jeder Baum auf der Grundlage von Informationen aus einem bestimmten Datensatz entwickelt wurde, besteht die Gefahr, dass das Modell zu sehr an den verwendeten Datensatz gebunden ist und die gemachten Prognosen nur begrenzte Gültigkeit besitzen und sich somit nicht auf größere Stichproben übertragen lassen. Um diese mögliche Einschränkung zu überprüfen, bietet SPSS AnswerTree drei Möglichkeiten, welche bei der Berechnung nicht auf die Genauigkeit der Prognosen sonder die Ungenauigkeit, bezeichnet als Risiko, zurück greifen. Es besteht die Wahl zwischen, Partitionierung zum Erstellen von Testdaten, der Resubstitution sowie der Kreuzvalidierung. AnswerTree gibt nach der Erstellung des endgültigen Baumes mit einer der angegebenen Validierungsmethoden eine Risikoschätzung in Form einer Fehlklassifikationstabelle an. Die Risikoübersicht sowie die Methoden sollen im Folgenden näher erläutert werden.

Die Partitionierung besteht darin, den Klassifikationsbaum an einer Reihe von Fällen, einer so genannten Lernstichprobe, zu entwickeln und den gefundenen Klassifikationsbaum dann anhand einer anderen Stichprobe zu validieren. Ein Beispiel hierfür wäre die Vorhersage des Ozonwertes in Los Angeles. So könnte man einen Klassifikationsbaum anhand der Werte in den Jahren 1972 bis 1975 entwickeln und dann überprüfen, ob er die Daten der Jahre 1976-77 richtig vorhersagen kann. Dieses Modell muss jedoch mit Vorsicht verwendet werden, da man nicht davon ausgehen kann, dass die beiden Stichproben unabhängig voneinander sind.

In aktuellen Problemen ist meistens nur eine Stichprobe vorhanden und es gestaltet sich schwierig, eine weitere, unabhängige Stichprobe aus der gleichen Grundgesamtheit in ähnlichem Umfang zu finden. Daher werden oftmals „interne“ Schätzungen vorgenommen, bei denen anhand der Stichprobendaten ein Klassifikationsbaum erstellt wird, in dem dann dieselben Fälle der gleichen Stichprobe nochmals klassifiziert werden. Die Formel hierfür wird folgendermaßen geschrieben:

$$R(d) = 1/N \ \Sigma \ X(d(x_n) \neq j). \qquad \text{(Gl. 12)}$$

Das Konstrukt d entspricht der Formel, die anhand der Lernstichprobe L errechnet wurde. Führt der Ausdruck in der Klammer zu einer Missklassifikation, so ergibt die Funktion X =1, anderenfalls X=0. Je höher die Anzahl an Fehlklassifikationen, desto schlechter ist das Modell.

Diese Art der Modelltestung, die auch als Resubstitution bezeichnet wird, führt jedoch zu einer Unterschätzung der Fehlklassifikation von R (d).

In einer weiteren Methode könnte man die zu untersuchende Stichprobe zufällig auf zwei Gruppen aufteilen und anhand von einer der beiden einen Klassifikationsbaum erstellen, der im zweiten Schritt an der anderen Stichprobengruppe getestet wird. Die Anzahl der falschen Klassifikationen N_2 aus der Lernstichprobe L_2 gäbe dann Aufschluss über die Güte meines Klassifikationsbaumes. Dieses Verfahren bezeichnet man auch als test sample Schätzung.

Die Berechnung erfolgt nach der folgenden Formel:

$$R^{ts}(d) = 1/N_2 \Sigma \; X(d(x_n) \neq j). \qquad \text{(Gl. 13)}$$

Auch in diesem Fall ist jedoch Vorsicht geboten, da unbedingt darauf zu achten ist, dass die Stichproben unabhängig voneinander sind, dennoch aus der gleichen Grundgesamtheit stammen. Letzteres kann am ehesten mit der Randomisierung gewährleistet werden. Ein weiterer Nachteil dieses Verfahrens liegt darin, dass die Stichprobe, mit welcher der Klassifikationsbaum erstellt wird, drastisch reduziert wird.

Bei größeren Stichproben besteht daher die Möglichkeit der Kreuzvalidierung. Nach diesem Verfahren wird die Stichprobe in mehrere gleichgroße Untergruppen aufgeteilt. Der Klassifikationsbaum könnte anhand von jeder Stichprobe erstellt werden. Anhand der Gesamtstichprobe wird sodann ein Klassifikationsbaum erstellt, der an den einzelnen Untergruppen untereinander validiert wird. Dieses Verfahren führt zu sehr zufriedenstellenden Vorhersagen, mit Werten, die simulierten Studien nahe kommen.

In allen Fällen kann R (d) dann als eine Wahrscheinlichkeit definiert werden, mit der das Konstrukt d eine neue Stichprobe, die aus einer gemeinsamen Grundgesamtheit gezogen wurde, falsch klassifizieren wird. Je kleiner die Fehlerklassifikation, desto besser ist das Modell.

Beurteilung der Klassifikationsanalyse

Klassifikationsbäume greifen den intuitiv einsichtigen Vorgang der Subgruppenbetrachtung auf, um ihn durch eine Formalisierung vor Willkür und Zufälligkeit zu schützen. Im Folgenden sollen seine Vor- und Nachteile erörtert werden.

Zunächst einmal stellt es ein sehr mächtiges und flexibles Hilfsmittel zur Klassifikation dar, welches automatisch und schrittweise Variablen selektiert und Komplexität reduziert. Auch die Bildung von Subgruppen, die dem intuitiven Vorgehen einer stratifizierten Analyse entspricht, zeichnet die Klassifikationsbäume aus und erlaubt auch eine Analyse sehr heterogener Gruppen. Ein weiterer Vorteil liegt in der übersichtlichen Darstellung und einfachen Interpretation der Ergebnisse, die auch Nicht-Fachleuten zugänglich sind. Klassifikationssysteme lassen sich ausserdem sehr leicht in der Praxis anwenden und ohne großen technischen Aufwand durchführbar. Sie können ohne spezielle Software und rechnerische Kenntnisse durchgeführt werden und auch fehlende Werte werden durch die Bildung einer eigenen Kategorie für diese problemlos mit berücksichtigt werden. Die CART-Analyse hat keine Voraussetzung und kann auch auf spezielle Subgruppen angewendet werden. Darüber hinaus lassen sie sich unmittelbar in Handlungsweisen oder Tests überführen. Es liefert ohne zusätzliche Mühe nicht nur eine Klassifikation, sondern auch eine Schätzung der Missklassifikation.

Nachteilig haben sich die Notwendigkeit einer großen Stichprobengröße und die unterschiedlichen Methoden zur Berechnung der Split- und Stop Regeln erwiesen, die zudem zu verschiedenen Ergebnissen führen können. Darüber hinaus besteht die Gefahr der Überinterpretation der Ergebnisse, insbesondere dann, wenn mit schwachen Split- und Stop-Regeln gearbeitet wird. Die Gestalt eines Klassifikationsbaumes ist im wesentlichen vom ersten Split abhängig, und damit bei ähnlichen Datensätzen nicht unbedingt identisch. Störvariablen können gerade bei kleineren Subgruppen Zusammenhänge suggerieren, die sich nicht validieren lassen. Schlussendlich stellen auch stetige Variablen bei der CART-Analyse ein Problem dar, da sie nur schwer in zwei Subgruppen aufgeteilt werden können.

Trotz der Nachteile sollte die Klassifikationsanalyse aufgrund ihrer vielen Vorteile und Lösung von Klassifikationsproblemen vermehrt eingesetzt und verbessert werden.

Literaturverzeichnis

Breimann, L., Friedmann, J.H., Ohlsen,R.A., Stone, C.J. (1993). *Classification and Regression Trees*. London: Chapmann &Hall

Bühl, A.. & Zöfel, P. (2002). *Erweiterte Datenanalyse mit SPSS, Statistik und Data Mining*. (Auflage); Wiesbaden: Westdeutscher Verlag.

Lefering, R.H.(1996). *Klassifikationsbäume- Ein multivariates Prognosemodell in der klinischen Anwendung und im Vergleich zur logistischen Regression* (Dissertation). Köln: Copy Team, Medizinische Fakultät der Universität zu Köln

Reiß, S & Moosbrugger, H. Prädiktoren von Studiendauer und Studienerfolg: Ergebnisse einer Absolventenbefragung im Diplom-Studiengang Psychologie der Johann Wolfgang Goethe Universität Frankfurt (WS 1995/96 – SS 2002).

Neuronale Netze

Augustin Kelava

Einleitung

In der Realität sind Wirkbeziehungen zwischen Variablen in der Regel sehr komplex. Diese Komplexität äußert sich in verschiedenen Aspekten; z.B.:

- in einer großen Anzahl von mit einander verknüpften Einflussfaktoren,
- in einer Vielzahl von Wechselwirkungen (seien es Wechselwirkungen im traditionellen statistischen Sinne oder gar reflexive Beziehungen),
- und nicht zuletzt in der Nicht-Linearität von Beziehungen zwischen den Variablen.

In vielen Fällen befindet sich der Forscher zudem in einem Stadium der Erkenntnisgewinnung, in dem er *keine begründeten Hypothesen über die Art der Ursache-Wirkungs-Beziehung* aufstellen kann. In seltenen Fällen ist er vielleicht noch (oder auch definitiv[4]) nicht daran interessiert, die Wirkungsbeziehungen aufzudecken, sondern vielmehr steht die *Genauigkeit der Vorhersage des Ergebnisses* im Zentrum der Beobachtung (bspw. bei der Vorhersage eines Kriteriums).

In solchen Fällen sind sog. Künstlich Neuronale Netze (im folgenden KNN oder NN) hilfreich, da der Anwender bei dieser Gruppe von Analyseverfahren nicht zwingend eine Vermutung (in Form einer Hypothese) über die Art der Zusammenhänge aufstellen muss.

Im Falle der KNN werden:

- weder kausale Beziehungen zwischen den Variablen angenommen
- noch sind Verknüpfungen zwingend linear anzunehmen.

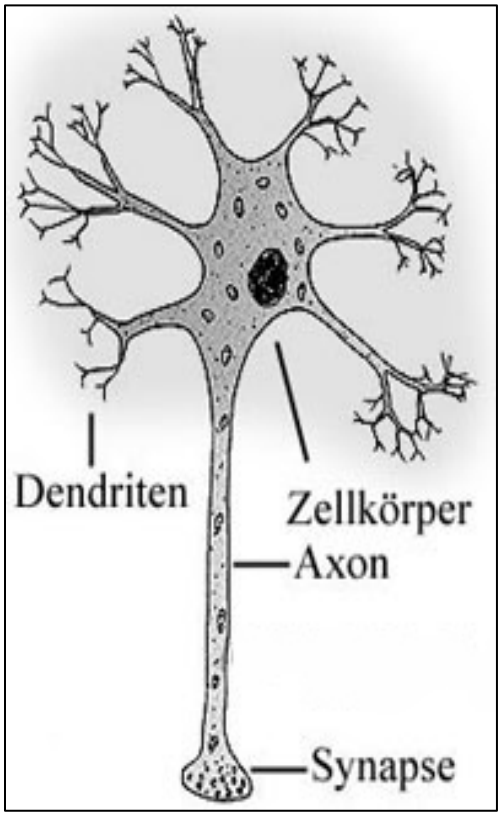
Künstlich Neuronale Netze ermitteln die Zusammenhänge zwischen Variablen selbständig durch einen *Lernprozess* und können dabei eine *Vielzahl von Variablen berücksichtigen* sowie *relativ komplexe mathematische Abbildungen der Zusammenhänge reproduzieren*. Sie dienen unter anderem dazu, Prognosen abzugeben (also Vorhersagen zu treffen) oder Klassifikationen vorzunehmen.

Doch um die Frage zu beantworten, was nun KNN sind, ist es zweckmäßig, die Parallelität der Funktionsweisen natürlicher Neurone und künstlichen Neuronen zu betrachten (siehe Abb. 1).

---

[4] Im Falle der ursprünglichen Anwendungsgebiete Künstl. Neuronaler Netze, der Erkennung von Mustern in der Datenverarbeitung oder seit geraumer Zeit bei der Datenkompression, ist man mehr am Ergebnis als an den Beziehungen der Variablen untereinander interessiert.

KNN wurden ursprünglich entwickelt, um Abläufe im Zentralen Nervensystem (ZNS) besser zu verstehen und/oder sie nachzubilden. Dabei dachte man sich künstliche Neurone ebenso wie natürliche Neurone aus folgenden Bestandteilen bestehend:



- *aus Dendriten*
- *aus einem Zellkörper (Soma)*
- *aus Axonen.*

Die Dendriten summieren dabei die Eingabe des Netzes (bzw. der vorangehenden Zellen) in die Zelle auf. Wird eine Nervenzelle über einen gewissen *Schwellenwert* S hinaus gereizt, erfolgt eine Erregung des Zellkörpers und die Weiterleitung der Erregung über die Axone bzw. die Ausgabe der Erregung der Zelle über den Synaptischen Spalt an die Dendriten nachfolgender Zellen

Abb. 1: Eine prototypische natürliche Nervenzelle

.

Die Stärke der Synapsen (die Stärke der Verbindung zweier Zellen) wird in KNN durch einen numerischen Wert, das *Verbindungsgewicht*, dargestellt. Wie auch bei natürlichen Neuronen erfolgt ein *Lernprozess* über die Veränderung der Verbindungsstärke.

Prinzipien Künstlich Neuronaler Netze

Aus welchen Komponenten bestehen nun KNN?

KNN bestehen im Wesentlichen aus *vier Hauptkomponenten*:

1. *Zellen*: Zellen (Neurone) kennzeichnen sich durch einen *Aktivierungszustand A* aus. Die *Aktivierungsfunktion* f_{acti} gibt an, wie sich ein neuer Aktivierungszustand $A_{neu,j}$ des Neurons j aus der alten Aktivierung $A_{alt,j}$ und der *Netzeingabe* $net_j(t)$ sowie dem *Schwellenwert* des Neurons j ergibt. Die *Ausgabefunktion* f_{out} bestimmt aus der Aktivierung die Ausgabe des Neurons.
2. *Verbindungsnetzwerk der Zellen*: Ein neuronales Netz kann als gerichteter, gewichteter Graph angesehen werden. Die Kanten stellen die Verbindungen zwischen den Neuronen dar. $w_{i,j}$ ist das Gewicht (weight) der Verbindung von Neuron i nach Neuron j (Vorsicht: umgekehrte Reihenfolge der Indices im Vgl. zur Pfadanalyse), die Matrix W aller Verbindungen heißt Gewichtsmatrix.

3. *Propagierungsfunktion*: Sie gibt an, wie sich die Netzeingabe eines Neurons $net_j(t)$ aus den Ausgaben der anderen Neuronen $o_{i,j}$ und den Verbindungsgewichten $w_{i,j}$ berechnet. Dabei handelt es sich um die gewichtete Summe der Ausgaben der Vorgängerzellen.

4. *Lernregel*: Sie ist ein Algorithmus, nach dem das Netz lernt, für eine vorgegebene Eingabe eine gewünschte Ausgabe zu produzieren. Durch die wiederholte Eingabe von Trainingsmustern wird die *Stärke der Verbindungen zwischen den Neuronen modifiziert*. Dabei wird versucht, den *Fehler* zwischen erwarteter und tatsächlicher Ausgabe des Netzes zu minimieren. Lernverfahren sind mit die interessanteste Komponente der neuronalen Netze (vgl. Rojas, 1993).

Komponenten einer Zelle

Formal kann man ein einzelnes *Neuron* folgendermaßen darstellen (nach Lippe, 2004):

Definition:

Ein künstliches Neuron ist ein Tupel $(x, w, f_{acti}, f_{out}, o)$ bestehend aus einem Eingabevektor $x = (x_1, ..., x_n)$, einem Gewichtsvektor $w = (w_1, ..., w_n)$, einer Aktivierungsfunktion f_{acti} mit f_{acti}: IR^n x IR^n ---> IR und einer Ausgabefunktion f_{out}, für die f_{out}: IR --> IR gilt. Dabei wird durch $f_{out}(f_{acti}(x, w)) = o$ der Ausgabewert des Neurons erzeugt, der an die nachfolgenden Neuronen über die Axone weitergeleitet wird.

Eine etwas anschaulichere Darstellung des Aufbaus eines künstlichen Neurons ergibt sich aus Abb. 2. Hier werden die Bestandteile eines Neurons nochmals dargestellt. Die Ausgaben o vorangehender Zellen werden aufgrund einer bestehenden Verbindungsstärke gewichtet, wie z.B. mit w_{gi}, zwischen der Zielzelle i und der Eingabezelle g. Die gewichteten Ausgaben der Vorgängerzellen werden anschließend anhand einer Propagierungsfunktion Σ zu einem Eingabewert net_j verdichtet (mathematisch betrachtet findet eine Zuordnung statt). Der Nettoeingabewert net_j bildet die Grundlage für einen Aktivierungszustand A_i, den die Zelle einnimmt. Der Aktivierungszustand ist zugleich die Variable für die Ausgabefunktion, die die Ausgabe o_i bzw. Erregung an ein nachfolgendes Neuron bestimmt.

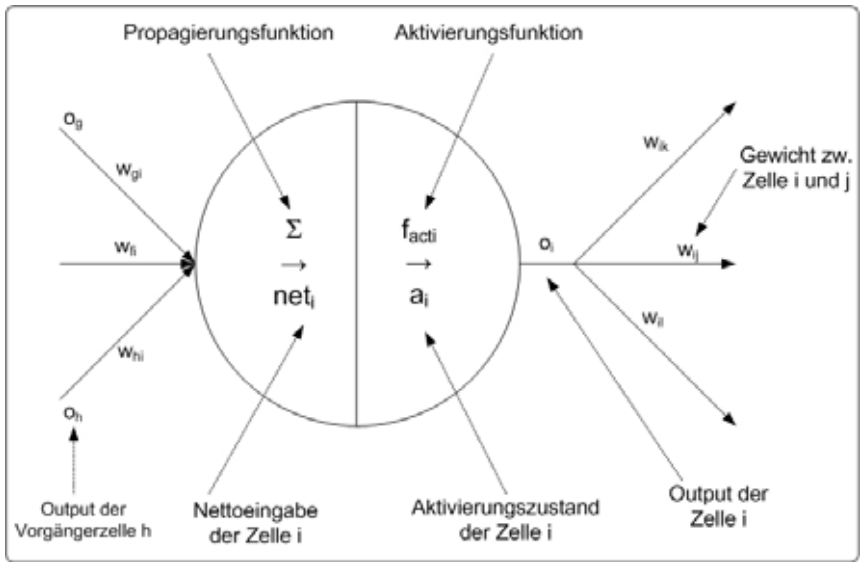


Abb. 2: Schematische Darstellung eines Neurons

Oftmals fasst man die Aktivierungs- und Ausgabefunktion zu einer *Transfer-Funktion* zusammen, dabei verwendet man die gewichtete Summe der Eingaben als Aktivierung und als Ausgabe:

$$f_{acti}(x, w) = \sum_{k=i}^{n} x_i w_i \qquad \text{(Gl. 1)}.$$

Es ist allerdings sinnvoller einem Neuron, analog zur natürlichen Nervenzelle, einen inneren *Aktivierungszustand* A_i zuzuordnen. Dieser hängt vom alten Zustand und der Veränderung der Aktivierungsfunktion ab und kann beispielsweise durch

$$A_{neu} = A_{alt} + f_{acti}(\vec{x}, \vec{w}) \qquad \text{(Gl. 2)}$$

definiert werden.

In der Realisierung eines Zustandes eines Neurons sind *viele Wertebereichte denkbar*, die je nach Anwendungsgebiet unterschiedlich beschränkt ausfallen können (z.B. kontinuierliche oder diskrete Verteilungen, unbeschränkte oder beschränkte Intervalle etc.)[5].

Bei kontinuierlichen Wertebereichen beschränken die meisten Modelle den Aktivierungszustand auf ein Intervall. Das liegt daran, dass es sich meistens um eine nichtlineare, häufig

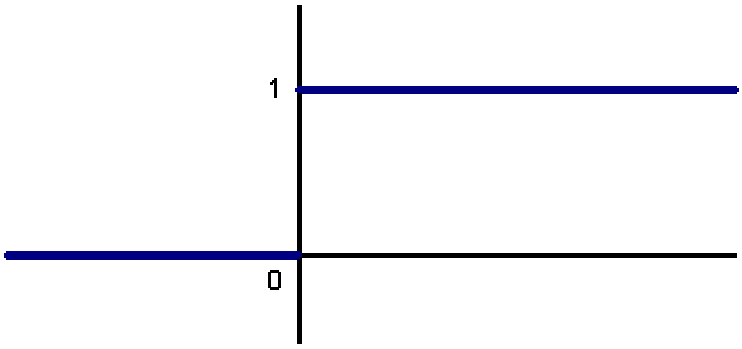
[5] Für Anwendungen aus der Informatik sind z.B. binäre Ausprägungen von Interesse, um Schaltungen zu simulieren. In der Psychologie, wo es eventuell um eine Prognose (ähnlich einer linearen Regression) geht, sind kontinuierliche uneingeschränkte Wertebereiche für gewisse Fragestellungen plausibel.

sigmoide, Aktivierungsfunktion und die Identität als Ausgabefunktion handelt. Dadurch wird die Ausgabe identisch mit der Aktivierung, und der Wertebereich der Aktivierungsfunktion gibt den Wertebereich des Aktivierungszustandes an.

Damit ein natürliches Neuron "feuert", also ein Aktionspotential ausgelöst wird, muss ein bestimmter Schwellenwert S überschritten werden. Auch für künstliche Neuronen gibt es eine *Schwellenwertfunktion*, sie könnte etwa lauten:

$$f_{out}(\sum_{k=1}^{n} x_k w_k) = \begin{cases} 1 : falls \sum_{k=1}^{n} x_k w_k > S \\ 0 : sonst \end{cases} \qquad \text{(Gl. 3)}$$

Der dazugehörige Funktionsgraph wäre wie folgt (Abb. 3):



Da diese Art der Ausgabefunktion nicht die Intensität von aufeinander folgenden Aktionspotentialen eines natürlichen Neurons simulieren kann, werden lineare Ausgabefunktionen verwendet.

Abb. 3: Graph einer einfachen Schwellenwertfunktion nach Gl. 3

Der zeitliche Abstand, in dem die Aktionspotentiale durch die natürliche Nervenzelle weitergereicht werden, ist allerdings nach unten beschränkt. Daher sollte auch im künstlichen Neuronenmodell eine beschränkte Ausgabefunktion Verwendung finden.

Diese Ausgabefunktionen lassen sich durch *semilineare Funktionen* der folgenden Art beschreiben:

$$f_{out}\left(\sum_{k=1}^{n} x_k w_k\right) = \begin{cases} 1 : falls \sum_{k=1}^{n} x_k w_k \geq \frac{1+a}{s} \\ s\left(\sum_{k=1}^{n} x_k w_k\right) - a : falls \frac{a}{s} \leq \sum_{k=1}^{n} x_k w_k < \frac{1+a}{s} \\ 0 : sonst \end{cases} \quad \text{(Gl. 4)}$$

Sinnvoller ist es aber, die Aktivierung beziehungsweise die Ausgabe durch glattere, also differenzierbare Funktionen zu beschreiben. Ein Beispiel für differenzierbare und beschränkte Funktionen sind die *s-förmigen oder sigmoiden Funktionen*, die in Abb. 5 unten dargestellt sind.

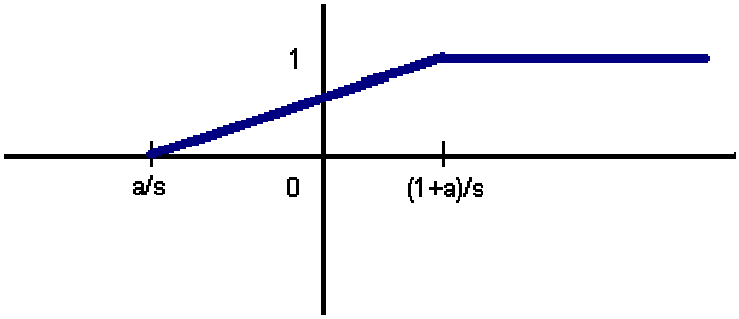


Abb. 4: Graph einer semilinearen Ausgabefunktion nach Gl. 4

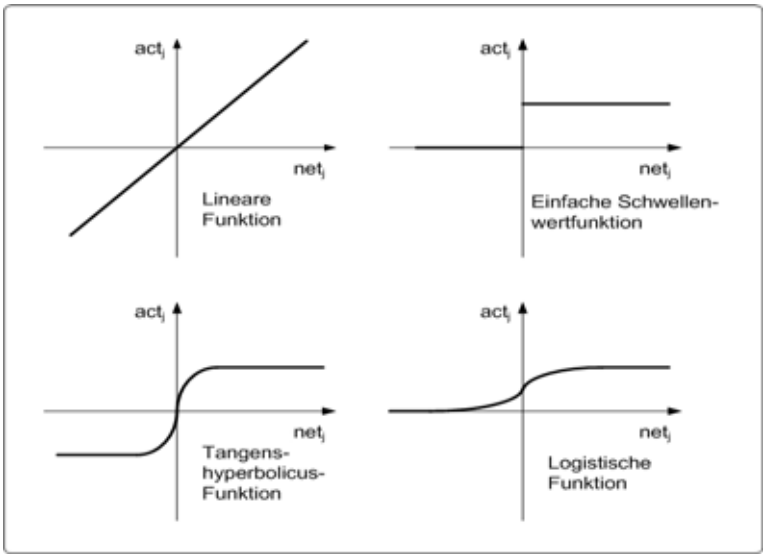


Abb.5: Beispiele für häufig verwendete Aktivierungsfunktionen

Das Verbindungsnetzwerk der Zellen / Die Topologie

Verbindet man nun mehrere Neuronen miteinander, so erhält man ein neuronales Netz.

Definition *(nach Lippe, 2004)*:

„Ein *Neuronales Netz* ist ein Paar (N, V) mit einer Menge N von Neuronen und einer Menge V von Verbindungen. Es besitzt die Struktur eines gerichteten Graphen, für den die folgenden Einschränkungen und Zusätze gelten:

1. Die Knoten des Graphen heißen Neuronen.
2. Die Kanten heißen Verbindungen.
3. Jedes Neuron kann eine beliebige Menge von Verbindungen empfangen, über die es seine Eingabe erhält.
4. Jedes Neuron kann genau eine Ausgabe über eine beliebige Menge von Verbindungen aussenden.
5. Das Neuronale Netz erhält aus Verbindungen, die der "Außenwelt" entspringen, Eingaben und gibt seine Ausgaben über in der "Außenwelt" endende Verbindungen ab.“

Ein Beispiel für ein einfaches Neuronales Netz, das der Definition entspricht sei im Folgenden abgebildet (Abb. 6):

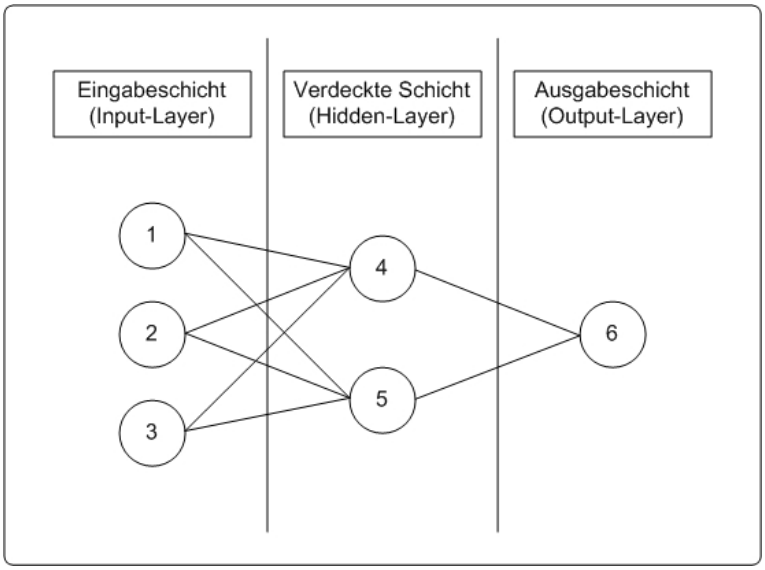


Abb. 6: Einfaches (2-schichtiges) Neuronales Netz mit einer Eingabe und einer Ausgabeschicht

In Abbildung 6 stellen die Neuronen 1-3 die Eingabeschicht dar. Das sind vorangehende Neuronen deren Ausgaben an Neuronen einer sog. verdeckten Schicht (hidden layer) weitergegeben werden. Diese Schicht kennzeichnet sich dadurch, dass ihre Modellierung

bzw. Annahme es erlaubt, komplexere auch nicht-lineare Zusammenhänge zu erfassen. I. d. R. beschränkt man sich auf max. 2-3 verdeckte Schichten, bis die „Erregung“ bzw. die Informationen der Eingabeschicht auf die Ausgabeschicht treffen.

Im Allgemeinen sind viele Ordnungsgesichtspunkte für Neuronale Netze denkbar. Ein Wesentlicher ist dabei die Art der Informationsverarbeitung. Man unterscheidet:

1. Netze ohne Rückkopplung (*feedforward-Netze*)
2. Netze mit Rückkopplungen (*rekurrente Netze*).

Zu 1. Bei *Netzen ohne Rückkopplungen* existiert kein Pfad, der von einem Neuron direkt oder über zwischengeschaltete Neuronen wieder zurück zu diesem Neuron führt. Daten werden also nur in eine Richtung weitergegeben (siehe Abb. 7). *Ebenenweise verbundene feedforward-Netze* sind in mehrere Schichten eingeteilt, wobei es nur Verbindungen von einer Schicht zur nächsten gibt (Teil A). Man spricht von vollständig verbundenen Netzen, falls jedes Neuron der Schicht U_i mit jedem Neuron der darauffolgenden Schicht U_{i+1} verbunden ist. *Allgemeine feedforward-Netze* besitzen dagegen auch sogenannte shortcut connections („Abkürzungen“), also Verbindungen zwischen Neuronen, die Ebenen überspringen (Teil B).

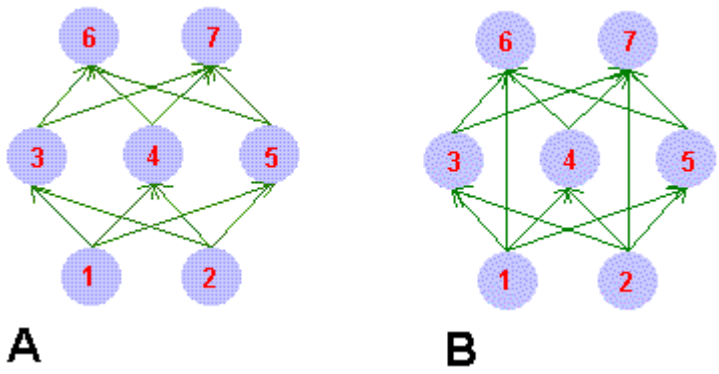


Abb. 7: Beispiele für Netze ohne Rückkopplungen

Zu 2. *Netze mit Rückkopplungen* unterteilt man meist folgendermaßen (s. Abb. 8):

- *Netze mit direkten Rückkopplungen (direct feedback)*
  Die Neuronen haben eine Verbindung von ihrer Ausgabe zurück zur Eingabe und können dadurch ihre eigene Aktivierung verstärken oder abschwächen. Diese Verbindungen bewirken oft, dass Neuronen die Grenzzustände ihrer Aktivierungen annehmen, weil sie sich selbst verstärken oder hemmen (Teil A).

- *Netze mit indirekten Rückkopplungen (indirect feedback)*
  Diese Netze besitzen Rückkopplungen von Neuronen höherer Ebenen zu Neuronen niederer Ebenen. Dadurch erreicht man eine Aufmerksamkeitssteuerung auf bestimmte Bereiche von Eingabeneuronen oder auf bestimmte Eingabemerkmale durch das Netz (Teil B).

- *Netze mit Rückkopplungen innerhalb einer Schicht (lateral feedback)*
  Netze mit Rückkopplungen innerhalb derselben Schicht werden oft für Aufgaben eingesetzt, bei denen nur ein Neuron einer Gruppe aktiv werden soll. Jedes Neuron hat dann hemmende Verbindungen zu den anderen Neuronen und oft noch eine aktivierende direkte Rückkopplung zu sich selbst. Das Neuron mit der stärksten Aktivierung, der „Gewinner", hemmt dann die anderen Neuronen. Daher heißt eine solche Topologie auch *winner-takes-all-* Netzwerk (Teil C).

- *vollständig verbundene Netze*
  Vollständig verbundene Netze haben Verbindungen zwischen allen Neuronen. Sie sind insbesondere als Hopfield- Netze bekannt geworden. Bei diesen muss allerdings auch die Verbindungsmatrix (Gewichtsmatrix) symmetrisch sein und die Diagonale darf nur Nullen enthalten (Teil D).

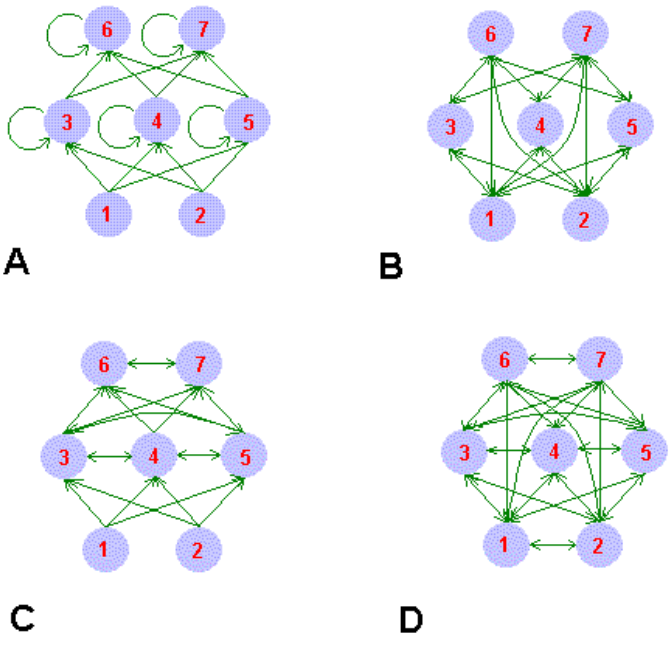


Abb. 8: Beispiele für Netze mit Rückkopplungen

Modellierung des Lernens

Ein neuronales Netz "lernt", indem es sich entsprechend einer fest vorgegebenen Vorschrift, der *Lernregel*, selbst modifiziert. Prinzipiell kann der Lernprozess bestehen aus:

- der Entwicklung neuer Verbindungen,
- dem Löschen existierender Verbindungen,
- der Modifikation der Verbindungsstärke (Veränderung der Gewichte)
- der Modifikation des Schwellenwertes
- der Modifikation der Aktivierungs- bzw. Ausgabefunktion
- der Entwicklung neuer Zellen
- dem Löschen bestehender Zellen.

Von diesen Möglichkeiten wird die dritte, also das Lernen durch Veränderung der Gewichte, am häufigsten verwendet (Auf dies soll in einem gesonderten Abschnitt eingegangen werden.). Erst in letzter Zeit haben Verfahren, die auch eine Veränderung der Topologie beinhalten an Bedeutung gewonnen.

Eine weitere Unterscheidungsmöglichkeit besteht in der Art des verwendeten Lernparadigmas. Hier lassen sich drei Arten unterscheiden:

1. Überwachtes Lernen (supervised learning)
Beim überwachten Lernen gibt ein externer „Lehrer“ dem Netz zu jeder Eingabe die korrekte Ausgabe oder die Differenz der tatsächlichen zur korrekten Ausgabe an. Anhand dieser Differenz wird dann das Netz über die Lernregel modifiziert. Diese Technik setzt allerdings voraus, dass Trainingsdaten exisitieren, die aus Paaren von Ein- und Ausgabedaten bestehen.
Ein typisches überwachtes Lernverfahren wie z.B. *Backpropagation* durchläuft für alle Paare von Ein- und Ausgabemustern folgende Schritte:

1. Das Eingabemuster wird dem Netz durch entsprechende Aktivierung der Eingabeneuronen präsentiert.
2. Die angelegte Eingabe läuft vorwärts durch das Netz. Dadurch wird ein Ausgabemuster für die aktuelle Eingabe erzeugt.
3. Tatsächliche und korrekte Ausgabe werden verglichen und die Differenz berechnet (Fehler).
4. Die Fehler laufen rückwärts von der Ausgabe- zur Eingabeschicht. Dabei werden die Verbindungsgewichte verändert, so dass der Fehler verringert wird.
5. Die Gewichte aller Neuronen werden um die vorher berechneten Werte verändert.

2. Bestärkendes Lernen (reinforcement learning)
Im Gegensatz zum überwachten Lernen wird dem Netz hier lediglich mitgeteilt, ob seine Ausgabe korrekt oder inkorrekt war. Das Netz erfährt nicht den exakten Wert des Unterschiedes.

3. Unüberwachtes Lernen (unsupervised learning)
Hierbei gibt es überhaupt keinen externen Lehrer, daher heißt dieses Lernparadigma auch *self-organized learning*. Das Netz versucht ohne Beeinflussung von außen die präsentierten Daten in Ähnlichkeitsklassen aufzuteilen.

Betrachtet man die Kombinationsmöglichkeiten verschiedener Lernregeln und Informationsflüsse, wird schnell deutlich, dass die KNN ein ganzes Methodenarsenal bereitstellen (Abb. 9).

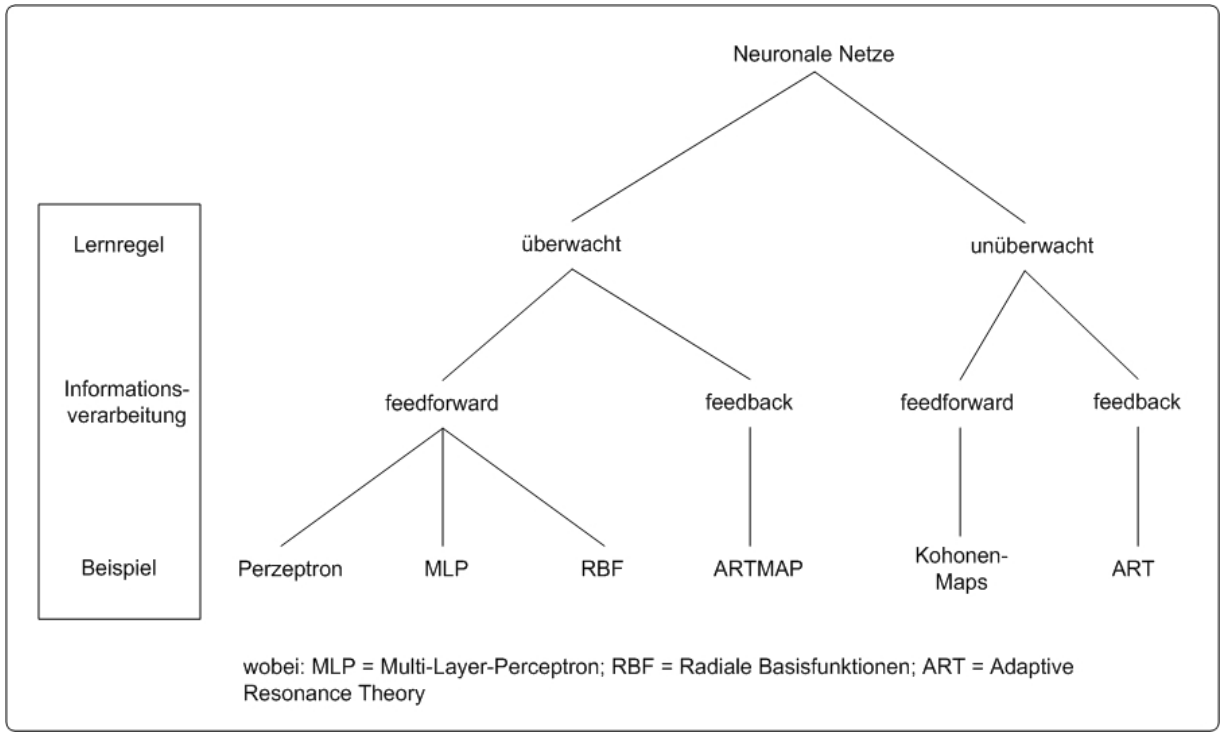


Abb. 9: Künstlich Neuronale Netze als Methodenarsenal; Unterscheidung und Klassifikation nach Lernregelparadigma und Informationsverarbeitungsrichtung (in Anlehnung an Backhaus et al., 2003)

Prinzipielle Vorgehensweise und ein anschauliches Kurzbeispiel

Beispiel in Anlehnung an Backhaus et al. (2003)

Man stelle sich vor, man sei daran interessiert den Kauf eines Produktes (z.B. einer Margarine) aufgrund dreier Eingangsvariablen vorherzusagen. Dies seinen:

- der Preis des Produktes in €
- das Geschlecht des Käufers
- das Gesundheitsbewusstsein des Käufers (anhand einer Ratingskala von 0 bis 10 erfasst).

Neben der Erfassung der Eingangsvariablen sei es gelungen, den potentiellen Käufern beim Kauf oder Nicht-Kauf der Margarine zu beobachten.

Ein Ausschnitt der entsprechenden Daten, mit denen das Neuronale Netz zwecks Vorhersage trainiert werden soll, sei nachfolgend abgebildet (s. Tab. 1):

Tabelle 1: Fiktives Beispiel für einen Trainingsdatensatz nach Backhaus et al. (2003)

	Geschlecht	**Preis**	**Gesundheitsbewusstsein**	**Kaufverhalten**
Person 1	m - 1	1,80 €	8	(1) Kauf
Person 2	w - 0	2,00 €	8	(0) Nichtkauf
Person 3	m - 1	1,50 €	9	(0) Nichtkauf
Person 4	w - 0	2,50 €	2	(1) Kauf
…	**…**	**…**	**…**	**…**

Für das vorliegende Beispiel wollen wir uns auf ein sog. Multi-Layer-Percetron (MLP) beschränken, das aus einer Eingabe-, einer verdeckten und einer Ausgabe-Schicht besteht (siehe Abb. 10; zu MLP-Netzen vgl. Zell, 2000).

Bevor der *erste Schritt* der Informationsverarbeitung in KNN durchgeführt werden kann, müssen die Gewichte zwischen den Neuronen (w_{ij}) zufällig auf von Null verschiedene Werte gesetzt werden (auch dies sei in Abb. 10 geschehen).

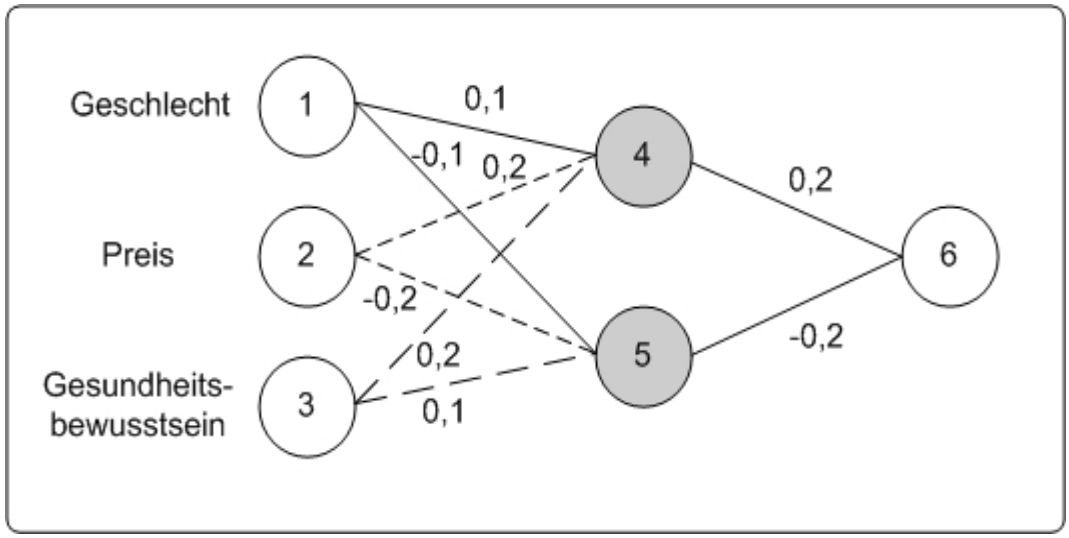


Abb. 10: Fiktives MLP mit Startwerten für das Kauf-Beispiel

Im *zweiten Schritt* ist es notwendig eine geeignete *Propagierungsfunktion* zu wählen, die die Ausgabewerte der vorgelagerten Schicht zu einem eindimensionalen Eingabewert für das nächste Neuron zusammenfasst.

Unter der Voraussetzung man habe eine einfache Summenwertfunktion gewählt, die eine gewichtete Summe bildet, würde sich für Person 1 der Nettoeingabewert in Neuron 4 und 5 wie folgt ergeben:

net_4: 0,1*1 + 0,2*1,8 + 0,2*8 = 2,06; net_5: -0,1*1 - 0,2*1,8 + 0,1*8 = 0,34

Wie in den vorangegangenen Abschnitten beschrieben, ist für jedes Neuron eine geeignete *Aktivierungsfunktion* f_{acti} zu wählen. Für unser Beispiel sei für $net_j > 0,5$ eine Aktivierung (als Schwellenwert) vorgesehen.

Damit würde für Person 1 sich folgendes ergeben:

$act_4 = 1$; $act_5 = 0$

$net_6 = 0,2 * 1 + (-0,2)*0 = 0,2$

$act_6 = 0$ (Nicht-Kauf ist vorhergesagt worden.)

Im *vierten Schritt* ist der Fehler zu bestimmen. Hierbei sei für Person 1 ein quadratischer Fehler von $(1-0)^2 = 1$ ermittelt worden, da ein Nichtkauf vorhergesagt worden ist, obwohl ein Kauf stattgefunden hat.

Dies verlangt im fünften Schritt eine Modifikation der Gewichte nach einem Lernalgorithmus, bevor im nächsten Schritt mit der Berechnung der Ausgabewerte für Person 2 fortgeschritten würde (iteratives Vorgehen). In einem späteren Abschnitt soll auf den Backpropagation-Algorithmus eingegangen werden. Dies ist ein Verfahren, das es erlaubt, die Veränderung der Gewichte effizient und strukturiert für mehrere Schichten vorzunehmen.

Die allgemeine Vorgehensweise

Die *allgemeine Vorgehensweise* umfasst dabei folgende Schritte:

1. Problemstrukturierung und Netztypauswahl
2. Festlegung der Netztopologie (siehe Abschnitt zur Topologie)
3. Bestimmung der genauen Informationsverarbeitung in den Neuronen (siehe Abschnitt zur Definition der Neuronen)
4. Trainieren des Netzes
5. Anwendung des trainierten Netzes

Zu 1.:

Bei der Problemstrukturierung und Netztypenauswahl ist es von Bedeutung, alle relevanten *Einflussfaktoren* auf die Zielvariablen zu bestimmen. Dies kann auf sehr vagen Hypothesen beruhen (dann handelt es sich um eine Strukturanalyse) oder aber der Zusammenhang ist gänzlich irrelevant. Eine Vorselektion der Variablen kann über bivariate Korrelationen erfolgen.

In Abhängigkeit des *Problemtyps* ist dann ein geeigneter *Netztyp* zu wählen. Ist zum Beispiel nach einer Prognose oder nach Ursache-Wirkungs-Zusammenhängen gefragt, bieten sich MLP oder RFB-Netze an. Ebenso eignen sich diese Netze zur Zuordnung zu bestehenden Gruppen. Bei Fragen der Klassifikation sind Kohonen-Maps, Hopfield-Netze und ART-Netze sinnvoll.

Zu 2.:

Nachdem der Netztyp festgelegt worden ist, ist die genaue Netztopologie zu wählen. Dabei ist die *Anzahl der verdeckten Schichten* zu definieren. In der Praxis haben sich bis zu max. 2 verdeckte Schichten etabliert (Dies liegt daran, dass die Mächtigkeit des Netzes nicht weiter zunimmt). Zudem ist festzulegen, wie hoch die Anzahl der Neuronen pro verdeckter Schicht ist. Dabei steigen der Trainings- und Rechen-Aufwand mit der Anzahl der Neuronen. Zudem besteht das „Problem des Übertrainierens“ bei zu hoher Anzahl der Neuronen. Das KNN kann zwar in einem solchen Falle sehr gut mit den Trainingsdaten umgehen, es hat aber die eigentliche Struktur des Problems nicht „erkannt“. Dies hat zur

Folge, dass eine Generalisierung auf neue Datensätze fehlschlägt und das Netz erneut trainiert und/oder grundlegend verändert werden muss.

Die *Struktur der Verbindungen zwischen den Neuronen* ist für eine gegebene Fragestellung geeignet zu wählen. In der Regel, sind alle Neuronen ebenenweise vollständig miteinander verbunden. Allerdings ist es sinnvoll auch Verbindungen wegzulassen oder direkte Verbindungen (short-cuts) zuzulassen, wenn theoretische Überlegungen das fordern. In Problematischen Fällen ist auch das Ausprobieren verschiedener Netztopologien zweckmäßig.

Zu 4.:

Im vierten Schritt ist das Netz mit Trainingsdaten zu trainieren. Ziel ist dabei, den Zusammenhang zwischen Input und Output möglichst optimal aufzudecken. Die Lernregel sieht meist eine Modifikation der Gewichte vor. Hierzu ist eine geeignete Methode zu wählen, wie etwa Backpropagation. Nachdem das Netz trainiert worden ist (dies entspricht i.d.R. 80% des Datensatzes), wird das Netz an einem Validierungsdatensatz angewendet bzw. trainiert. Dabei wird der Lernprozess optimiert (s. Abb. 11). Die Gewichte werden anhand des Validierungsdatensatzes nochmals verändert, so dass sich die eigentliche Struktur des Problems und nicht ausschließlich die Eigenschaften des Trainingsdatensatzes herauskristallisiert. Dabei ist auf eine optimale Anzahl von Lernschritten zu achten, damit die beiden Datensätze eine optimale Anpassung des Netzes bewirken.

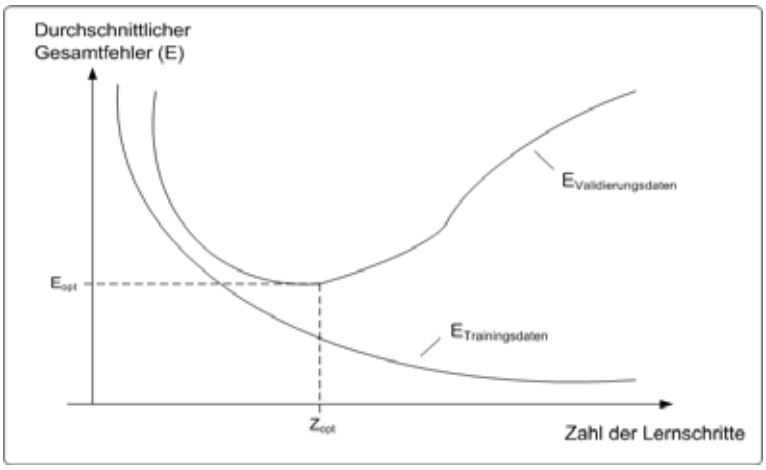


Abb. 11: Durchschnittlicher Gesamtfehler der Trainings- und Validierungsdatensätze in Abhängigkeit von der Anzahl der Lernschritte.

Zu 5.:

In letzten Schritt wird dann das KNN endgültig auf einen Testdatensatz angewendet, wobei aus den Eingabedaten Ausgabedaten berechnet werden. Die Gewichte werden bei in diesem Schritt eingefroren und nicht weiter verändert. Eine Fehleranalyse ist dabei erst dann möglich, wenn das zu prognostizierende Ereignis eingetreten ist. Das KNN ist so lange anwendbar, wie sich die Problemsituation nicht verändert.

Der Backpropagation-Algorithmus

Bereits der Name „Backpropagation-Algorithmus (McClelland & Rumelhart, 1986)“ sagt aus, dass die Modifikation der Gewichte rückwärtsgerichtet erfolgt, d.h. von der Ausgabe- hin zur Eingabeschicht. Dabei ermittelt die Fehlerfunktion den Fehler zwischen Soll-Ausgabe und der Netz-Ausgabe. Da die Fehlerfunktion an der Ausgabeschicht ansetzt, kann nur der Beitrag der Verbindung, der direkt zur Ausgabeschicht führt, unmittelbar berechnet werden. Für die anderen Gewichte wird eine Fortpflanzung (propagation) des Fehlers von der Ausgabeschicht zur Eingabeschicht unterstellt. Dementsprechend erfolgt auch die Veränderung der Gewichte rückwärtsgerichtet (back). Im Folgenden wird am Beispiel eines mehrschichtigen KNN die Informationsverarbeitung im sog. Multi-Layer-Perzeptron und das Training mit dem BP-Algorithmus dargestellt, wobei man *neun Ablaufschritte* unterscheidet:

Schritt 1: Initialisierung der Startgewichte: zufällige Ausgangswerte für alle w_{ij}

Für alle definierten Verbindungen w_{ij} zwischen zwei den Neuronen der verschiedenen Schichten werden zufällige Ausgangswerte gewählt.

Schritt 2: Berechnung der Ausgabewerte

Im vorwärtsgerichteten Schritt werden einzeln für jedes Datenmuster p die Netzausgaben berechnet.

Schritt 3: Berechnung des Netzfehlers

An der Ausgabeschicht wird der Fehler für das Datenmuster p durch die Fehlerfunktion E_p berechnet. Allgemein misst die Fehlerfunktion die Abweichung zw. den berechneten und den erhobenen Ausgabewerten für ein einzelnes Muster p. Allerdings ist nicht der Fehler eines Musters zu minimieren, sondern der durchschnittliche Fehler über alle Datenmuster (der sog. „mean square error“, MSE; vgl. Lippe, 2004). Der Backpropagation-Algorithmus versucht, den durchschnittlichen Gesamtfehler zu minimieren, indem er die Gewichte einzeln für jeden Datensatz p ändert. Das arithmetische Mittel aller Gewichtsänderungen ist eine Schätzung der wahren Änderung der Gewichte, die nötig wäre, um den durchschnittlichen Gesamtfehler zu minimieren (dies geschieht durch partielle Gradientenbestimmung).

In der Regel wird die quadratische Fehlerfunktion

$$E = \sum_{i=1}^{a} (o_i - t_i)^2 \qquad \text{(Gl. 5)}$$

Verwendet, wobei a die Anzahl der Neuronen in der Ausgabeschicht angibt. Für jedes Neuron berechnet die Fehlerfunktion die Differenz zwischen dem berechneten Ausgabewert o und dem empirisch ermittelten Ausgabewert t. Für Klassifizierungsaufgaben bspw. lassen

sich auch andere Fehlerfunktionen verwenden, etwa um Wahrscheinlichkeiten zu berücksichtigen.

$$E = [t \cdot o + (1-t) \cdot \ln(1-0)] \qquad \text{(Gl. 6)}$$

Außerdem sind Fehlerfunktionen denkbar, die die Kosten für die Fehler zw. Soll-Ausgabe und der berechneten Ausgabewerte erzeugen.

Schritt 4: Ermittlung der Suchrichtung

Die Gewichte sollen so verändert werden, dass der Fehler minimiert wird. Wird das Netztraining als Optimierungsproblem betrachtet, so kann/wird die Fehlerfunktion als Zielfunktion betrachtet. Für jedes Gewicht im Netz lässt sich mittels des *Gradientenverfahrens* die Richtung der Gewichtsänderung berechnen, die den Fehler am stärksten verringert. Der Gradient der Fehlerfunktion zeigt in die Richtung der steilsten Steigung an der aktuelle Stelle. Um den Gradienten auch für Verbindungen zwischen vorgelagerten Schichten berechnen zu können, wird das Konstrukt „*Fehlersignal*" eingeführt, das sich unmittelbar aus der mathematischen Herleitung des BP-Algorithmus (vgl. Zell, 2000 und Lippe 2004) ergibt. Die Fehlersignale werden rekursiv von der Ausgabeschicht ausgehend berechnet und erlauben die Ermittlung des Beitrages eines vorgelagerten Gewichtes für das Zustandekommen des Netzfehlers an der Ausgabeschicht.

Schritt 5: Bestimmung der Schrittweite bzw. Lernrate

Die sog. Schrittweite gibt wie bei vielen anderen Verfahren an, wie stark die Änderung der Gewichte in Richtung der steilsten Steigung erfolgen soll. Da sich in der Gewichtsänderung der Lernprozess abbildet, wird die Schrittweite auch als Lernrate η bezeichnet. Die Lernrate η hat einen Einfluss darauf, inwieweit der Lernalgorithmus konvergiert, d.h. ein Minimum der Fehlerfunktion des Trainingsdatensatzes finden kann. Je kleiner die Lernrate ist, umso mehr Lernschritte sind notwendig, bis ein Minimum der Fehlerfunktion erreicht wird. Die Änderung der Gewichte in einem Lernschritt ist geringer, d.h., der Algorithmus macht sehr kleine Schritte auf der Fehlerfläche der Fehlerfunktion. Dabei steigt natürlich der Rechenaufwand gerade bei komplexen Modellen. Die Schrittlänge sollte aber auch nicht zu groß gewählt werden, weil sonst die Gefahr besteht ein Minimum zu überspringen. Die besten Ergebnisse liefern Verfahren bzw. Implementierungen, die die Lernrate im Laufe des Lernprozesses verringern.

Schritt 6: Änderung der Gewichte

Die Änderung der Gewichte w_{ij} erfolgt nach folgendem Berechnungsschema, das auf alle Gewichte w_{ij} angewendet wird:

$$\Delta w_{ij} = \eta \cdot \delta_{pj} \cdot net_{pi} \qquad \text{(Gl. 7)}$$

Dabei bezeichnet Δw_{ij} die Gewichtsänderung, n die Lernrate, δ_{pj} ist das Fehlersignal und net_{pi} die Eingabe von Neuron i.

Schritt 7: Berechnung der neuen Netzausgabewerte

Durch die Verwendung der neuen Gewichte erhält man die Ausgabewerte für das folgende Datenmuster. Analog zu Schritt 3 wird dann wiederum der Fehler an der Ausgabeschicht berechnet.

Schritt 8: Berechnung des neuen Netzfehlers

Analog zu Schritt 3 wird der Netzfehler für die neuen Gewichte der Datenmuster berechnet.

Schritt 9: Überprüfung der Abbruchkriterien

Der BP-Algorithmus wird beendet, wenn die zuvor definierten Abbruchkriterien erfüllt sind, andernfalls wird mit Schritt 4 fortgefahren. Kann durch den BP-Algorithmus allein kein befriedigendes Ergebnis erzielt werden, so sind entweder die Neuronen anders zu definieren oder die Netztopologie zu ändern. Darüber hinaus kann das Abbruchkriterium oder die Lernrate verändert werden.

Problemfelder bei der Anwendung des BP-Algorithmus

Der Backpropagation-Algorithmus ist nicht frei von Problemen (für einen Überblick: s. Lippe, 2004). In Abb. 12 sei für ein Gewicht w_{ij} beispielhaft dargestellt, welche Probleme im Zusammenhang bei der Minimierung der Fehlerfunktion auftreten können.

Teilweise liegen die Probleme darin begründet, dass der BP-Algorithmus als Gradientenverfahren nur seine unmittelbare Umgebung und nicht die gesamte Fehlerfläche berücksichtigen kann. Der BP-Algorithmus berechnet ein lokales Minimum. Er kann aber nicht sicherstellen, ein global Minimum gefunden zu haben (links oben). Es kann auch keine aussage darüber getroffen werden, wie groß der Unterschied zw. lokalem und globalem Minimum ist. Möglicherweise ist ein lokales Minimum dennoch eine gute Näherung an das globale Minimum.

Ein ähnliches Problem liegt bei sog. flachen Plateaus vor. Auch hier ist es möglich, dass der Lösungsalgorithmus nicht das globale Minimum identifiziert, weil ein Abbruchkriterium keine bedeutsame Veränderung des Fehlers nach einer hohen Anzahl von Lernschritten mehr feststellt (rechts oben).

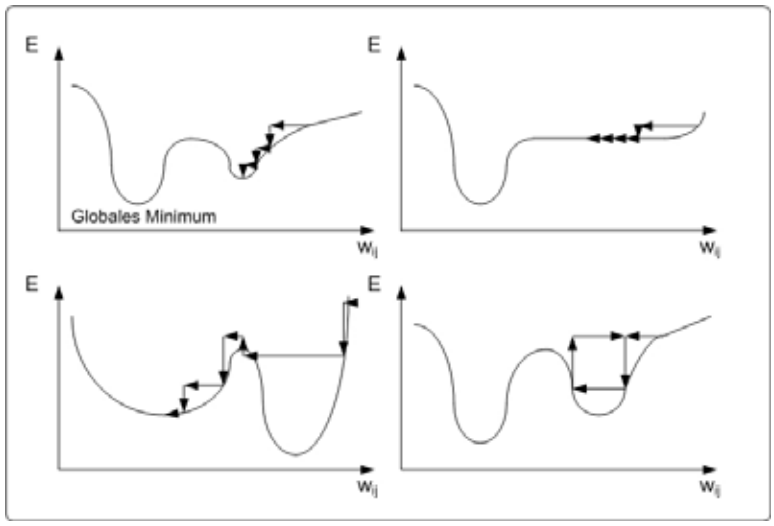


Abb. 12: Einige häufige Konvergenzpropleme

Wird eine größere Lernrate gewählt, sinkt die Gefahr, dass der BP-Algorithmus ein suboptimales lokales Minimum der Fehlerfunktion identifiziert oder in einem flachen Plateau hängen bleibt. Es kann aber auch vorkommen, dass der BP-Algorithmus gute Minima verlässt (links unten) und stattdessen ein suboptimales Minimum findet. Dies kommt in besonders engen Tälern vor.

Schließlich kann der Algorithmus in steilen „Schluchten“ der Fehlerfunktion auch oszillieren. Ist der Gradient am Rande einer Schlucht sehr groß, kann er an die andere Seite der Schlucht springen. Hat die Schlucht auf der anderen Seite dieselbe Steigung, wird er wieder zurückspringen (unten rechts).

Literaturverzeichnis

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2003). *Multivariate Analysemethoden* (10. Auflage). Berlin: Springer.

Lippe (2004). *Einführung in die Neuronale Netze* [available 10.03.2005: http://wwwmath.uni-muenster.de/SoftComputing/lehre/material/wwwnnscript/startseite.html].

McClelland, J.L., & Rumelhart, D.E. (1986). *Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Rojas, R. (1993). *Theorie der neuronalen Netze - Eine systematische Einführung.* Berlin: Springer.

Rumelhart, D.E., & McClelland, J.L. (1996). *Parallel Distributed Processing*: Exploration in the Micorstructure of Cognition. (Vol. 1). Cambridge, MA: MIT Press.

Zell, A. (2000). *Simulation Neuronaler Netze.* (3., unveränderter Nachdruck). Bonn: Addison-Wesley.

Messung von Studienerfolg über Studiennoten und Studiendauer

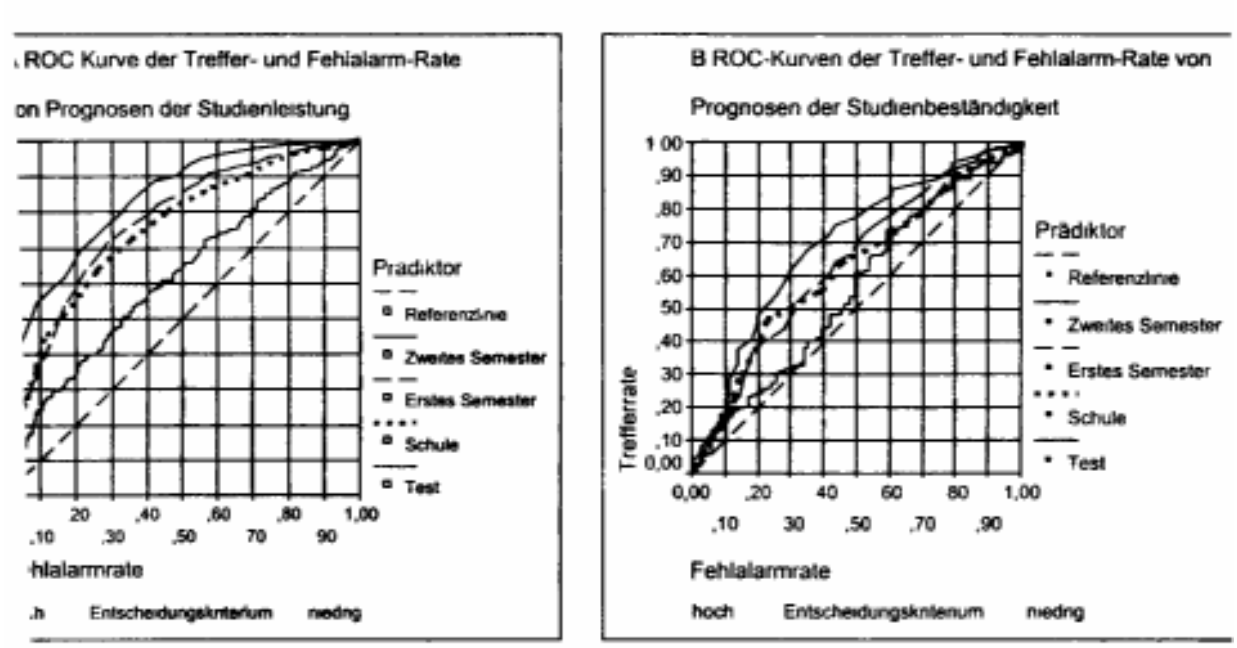
Birgit Menzel

Um ein geeignetes Verfahren zur Studierendenauswahl zu finden, muss man zunächst überlegen wie man den Studienerfolg am besten operationalisieren kann. Dafür gibt es verschiedene Ansätze. Wie im Folgenden dargestellt, kann man den Studienerfolg über die Studiennoten oder die Studiendauer erheben. Es ist aus ökonomischen Gründen für die Universität und die Studierenden sinnvoll eine geeignete Studierendenauswahl durchzuführen, da die Auswahl von zügig studierenden und guten Absolventen eine Kostenersparnis, für die Hochschule, ein gutes Ansehen und eine gute Ausbildung für einen schnellen Berufseintritt gewährleistet.

Studienerfolg über Studiennoten

In der Untersuchung „Erste Prüfungen – weiterer Studienerfolg“, die von H. Brandstätter und A. Farthofer 2004 an der Universität von Linz durchgeführt wurde, wurde die Vorhersagekraft von Noten vor und während des ersten Studienjahres für den weiteren Studienerfolg betrachtet. Hierfür wurden die Noten der Hauptfächer des letzten Schulzeugnisses von 789 Schülern, deren Leistung in einem studienfachspezifischen Test (SFT), der Notendurchschnitt des ersten und des zweiten Semesters verwendet. Die Schulnoten wurden, da die Befragung im Rahmen einer Studienberatung vier bis sechs Monate vor der Reifeprüfung stattfand, aus dem letzten Schulzeugnis entnommen. Außerdem wurden neben einigen motivationalen Fragen, der 16PA zur Beantwortung vorgelegt. Diese Persönlichkeits–Adjektive-Skala basiert auf den 16PF-Primärskalen, die aus zwei Listen mit 16 bipolaren Eigenschaftspaaren bestehen. Für diese Studie wurden die Primärfaktoren G und Q3 (Regelbewusstsein und Perfektionismus) verwendet. Ein weiterer wichtiger Punkt war eine Beurteilung der Aussage „Ich lege großen Wert auf gute Noten“ auf einer neunstufigen Skala. Die Leistungen des ersten und zweiten Semesters wurden mit dem Notendurchschnitt der Scheine in dem jeweiligen Semester ermittelt. Ferner wurde versucht, eine Vorhersage für die Wahrscheinlichkeit eines Abbruchs bzw. eines Abschlusses des Studiums zu machen. Zugrundegelegt wurde der Methode das Modell zur Signalentdeckung. Dabei wird einer Person entweder nur ein Rauschen oder ein Signal über dem Rauschen dargeboten. Die Person muss dann entscheiden, ob sie ein Signal wahrgenommen hat oder nicht. Dabei kann es zu Treffern (ein Signal wird angezeigt, wenn tatsächlich eines vorhanden ist) und zu Fehlalarmen (ein Signal wird angezeigt, obwohl keines dargeboten wurde) kommen. Die Entscheidungsfindung beruht auf mehreren Faktoren: die objektive Stärke des Reizes, die Sensibilität des Beurteilers und die

Reaktionstendenz. Letztere wird von der subjektiv wahrgenommenen Reizstärke und der Einschätzung der Konsequenzen eines Fehlalarms oder Übersehen eines Signals beeinflusst. Wendet man nun dieses Modell auf die Frage nach Studienerfolg an, stellt der Studienabbruch das Signal dar und die prognostische Validität der Schulnoten die objektive Reizstärke des Signals. Damit lässt sich eine ROC-Kurve (ROC: im SPSS = Receiver Operating Characteristic) mit dazugehöriger Referenzlinie in eine Grafik zeichnen. Sie veranschaulicht die Wahrscheinlichkeit für einen Studienabbruch bei einer bestimmten



Notenklasse.

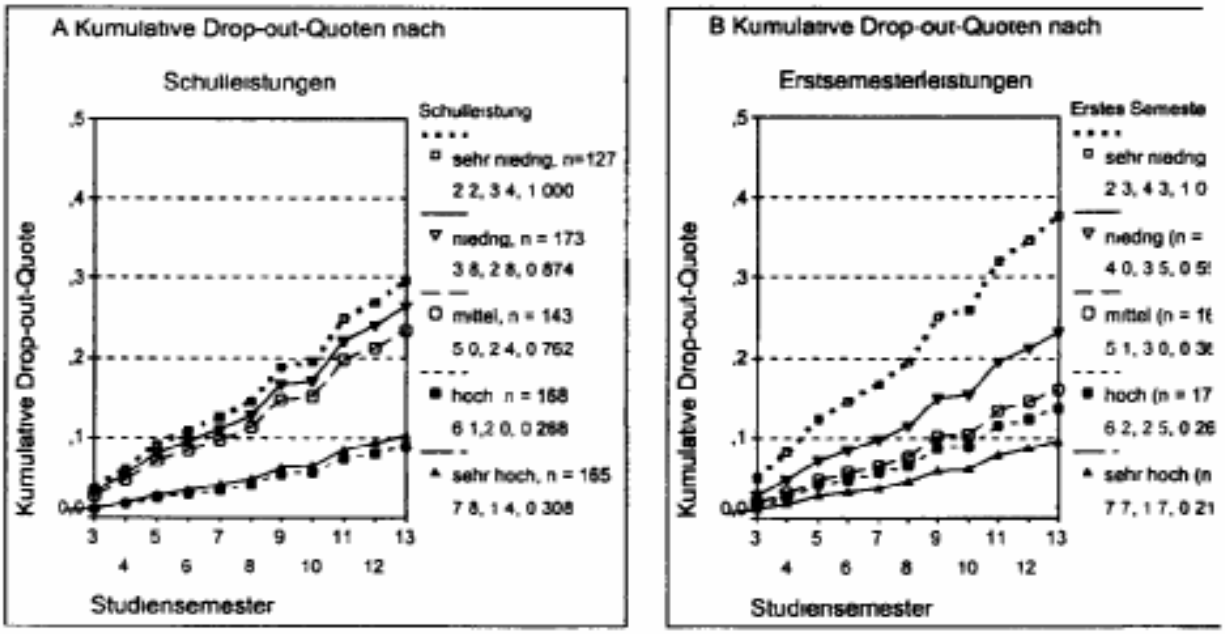
Abb. 1: Treffer- und Fehlalarmrate der Prognose überdurchschnittlicher Studienleistung (A) und der Studienbeständigkeit (B) ab dem dritten Semester aufgrund von Schulleistung, Testleistung, Erst- und Zweitsemesterleistung[6]

Ist nun die Wahrscheinlichkeit für einen Studienabbruch hoch (starkes Signal) bei einer mittleren Validität der Schulnoten (Rauschen) wird ein Abbruch auch gut prognostizier. Die Entdeckungsrate ist hoch. Dieses Modell hat den Vorteil , dass in einem binären Kriterium die Validität verschiedener Prädiktoren anschaulich dargestellt werden kann, und der Erkennbarkeit, ob die Trennschärfe über die Skalenerstreckung des Prädiktors homogen ist.

Die Ergebnisse der Untersuchung haben gezeigt, dass die Schulleistungen signifikant höher mit den Erstsemesterleistungen korrelieren als die Testleistungen im FST. Beide Prädiktoren zusammen können optimal gewichtet 29% der Varianz des Studienerfolgs im

[6] Aus Brandstätter, H. & Farthofer, A. (2003). Erste Prüfung – weiterer Studienerfolg aus Psychologie in Erziehung und Unterricht

ersten Semester erklären. Nach regressionsanalytischer Betrachtung ergibt sich für die Zweitsemesterleistung die besten Korrelationen mit dem weiteren Studienerfolg. Die vier Prädiktoren Test-, Schul-, Erst- und Zweitsemesterleistung erhöhen jeweils signifikant den Anteil erklärter Varianz des weiteren Studienerfolgs bis zu 48%. Die Testleistungen alleine sind kein guter Prädiktor für den Studienerfolg oder die dazu nötige Motivation, da der Test eine besondere Situation darstellt und die Ergebnisse auch nicht – im Gegensatz zu den Schulleistungen, die den normalen Zustand repräsentieren – mit den Persönlichkeitsvariablen Regelbewusstsein und Perfektionismus korrelieren. Die



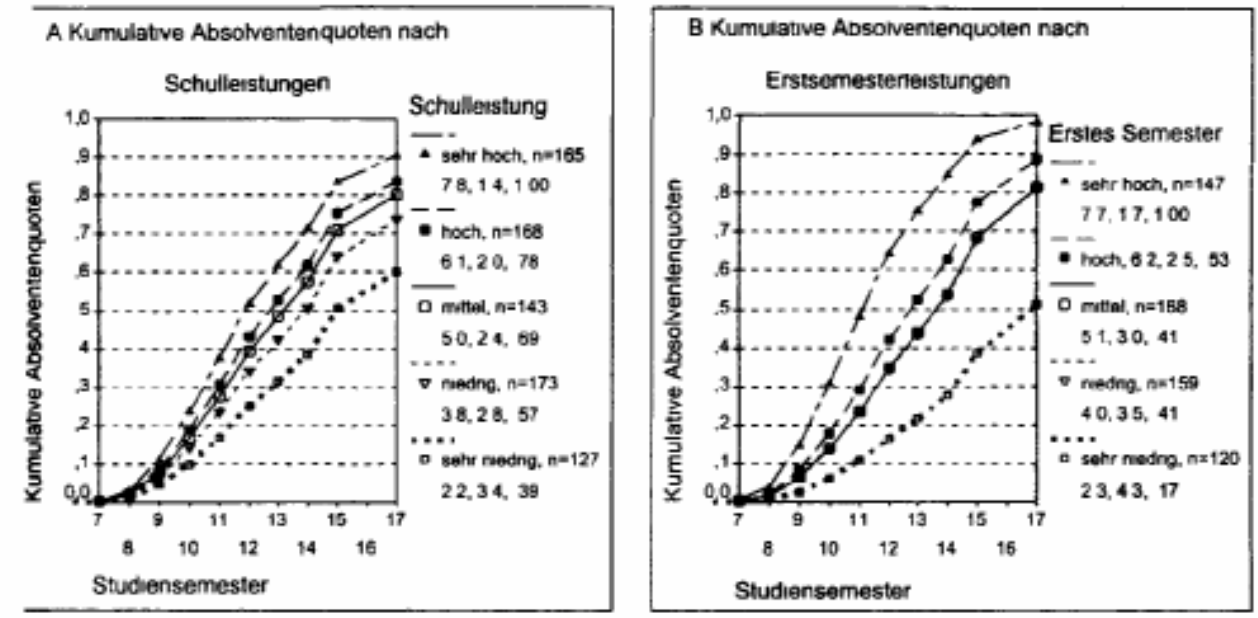
zufallskritische Prüfung ergab eine Signifikanz für $p < .001$, $p < .05$ und $p < .10$ einseitig (nach Bortz).

Abb. 2: Dropout-Quoten im Semesterverlauf (ab drittem Semester) in fünf nach Schul- bzw. Erstsemesterleistung geordneten Gruppen. Die Zahlen unterhalb der Gruppenkategorie bezeichnen die Schul- bzw. Erstsemesterleistung (erste Zahl Staninewert, zweite Zahl Notendurchschnitt) sowie das Verhältnis der Hasardrate der betreffenden Leistungsgruppe (der semesterspezifischen Dropout-Wahrscheinlichkeit, bezogen auf jene Studierenden, bei denen das Ereignis noch eintreten konnte) zur Hasardrate der Referenzgruppe (hier die Gruppe ‚Schulleistung sehr niedrig').[7]

Die Frage nach der Wahrscheinlichkeit für ein erfolgreiches Studium kann am besten durch die Leistungen im zweiten Semester (mit 34% Sicherheit) vorhergesagt werden, wobei sich

[7] Aus Brandstätter, H. & Farthofer, A. (2003). Erste Prüfung – weiterer Studienerfolg aus Psychologie in Erziehung und Unterricht

die Vorhersagekraft der Schulnoten nicht wesentlich von der der Erstsemesterleistung unterscheidet (ca. 25%). Nur die Testleistung (mit 12%) eignet sich nicht zur Vorhersage. Für die Dropout- oder Absolventenquoten (Dropout = Studiumsabbruch ⬭ Kriterium: zwei Semester ohne Leistungsnachweise; Studienortswechsel ausgeschlossen) lassen sich flachere ROC-Kurven verzeichnen, jedoch ist auch hier die Leistung des zweiten



Studiensemesters der beste Prädiktor. Um die Testleistung und die Noten besser miteinander vergleichen zu können, werden die Noten umgepolt (gute Leistung = hoher Wert ⬭ Standard-Nenner-Skala (Stanine)).

b. 3: Absolventenquoten im Semesterverlauf (ab drittem Semester) in fünf nach Schul- bzw. Erstsemesterleistung geordneten Gruppen. Die Zahlen unterhalb der Gruppenkategorie bezeichnen die Schul- bzw. Erstsemesterleistung (erste Zahl Staninewert, zweite Zahl Notendurchschnitt) sowie das Verhältnis der Hasardrate der betreffenden Leistungsgruppe (der semesterspezifischen Abschlusswahrscheinlichkeit, bezogen auf jene Studierenden, bei denen das Ereignis noch eintreten konnte) zur Hasardrate der Referenzgruppe (hier die Gruppe ‚Schulleistung sehr hoch').[8]

Es zeigt sich weiter, dass das Dropout-Risiko für leistungsstarke Studierende signifikant niedriger ist als für leistungsschwache, nämlich gleichgültig, ob mit dem Prädiktor Schulleistung oder Erstsemesterleistung gemessen wurde, ist sie ca. dreimal so hoch bei leistungsschwachen als bei leistungsstarken Studierenden. Erstaunlicherweise sind die

[8] Aus Brandstätter, H. & Farthofer, A. (2003). Erste Prüfung – weiterer Studienerfolg aus Psychologie in Erziehung und Unterricht

Dropout-Quoten in der Studie wesentlich niedriger als in den amtlichen Statistiken, was auf den FST zurückgeführt werden kann. Die Dropout-Quote wurde um fast ein Drittel gesenkt. Denn dieser Test, der studienfachspezifische Aufgaben enthält, macht die Studiumsanwärter darauf aufmerksam, welche Anforderungen auf sie zukommen werden. Damit führt er zu einer realistischeren persönlichen und Studienfach betreffenden Einschätzung. Allerdings ist der Notendurchschnitt im Studium besser vorherzusagen als der eventuelle Dropout. Denn den Dropout beeinflussen viele Faktoren, unter anderem auch der Notendurchschnitt.

Auch andere Ergebnisse wie die von H. Schaeper und K. Minks vorgelegte Arbeit „Studiendauer- eine empirische Analyse der Determinanten und Auswirkungen auf den Berufseintritt“ (1997) zeigen einen deutlichen positiven Zusammenhang sowohl zwischen dem Notendurchschnitt des Abschlusszeugnisses und der Examensnote als auch zwischen dem Notendurchschnitt im Abschlusszeugnis und der Studiendauer. Dies gilt vor allem in den technischen, natur- und wirtschaftswissenschaftlichen Fächern. Gute ehemalige Schüler haben auch gute Examensnoten und studieren durchschnittlich kürzer als schlechte. Ermittelt wurden die Erkenntnisse mittels der jeweiligen Prozentangaben von über- bis unterdurchschnittlich guten Schul- und Examensnoten und über- bis unterdurchschnittlich langen Studierzeiten. Anschließend wurde ein t-Test zur Signifikanzprüfung verwendet.

Studienerfolg über Studiendauer

Allgemeine Determinanten der Studiendauer

Die Dauer eines Studiums unterscheidet sich sowohl von Fach zu Fach als auch von Studienort zu Studienort. Im Folgenden werden die Faktoren dafür versucht aufzuzeigen. In der bereits oben erwähnten von H. Schaeper und K. Minks (1997) vorgelegten Arbeit, werden die Determinanten für die Studiendauer in individuelle und institutionelle Faktoren unterteilt und auf ihren Einfluss und Wechselwirkungen untersucht. Dabei konnte kein eindeutiges Übergewicht einer Faktorengruppe bei der Einflussnahme festgestellt werden.

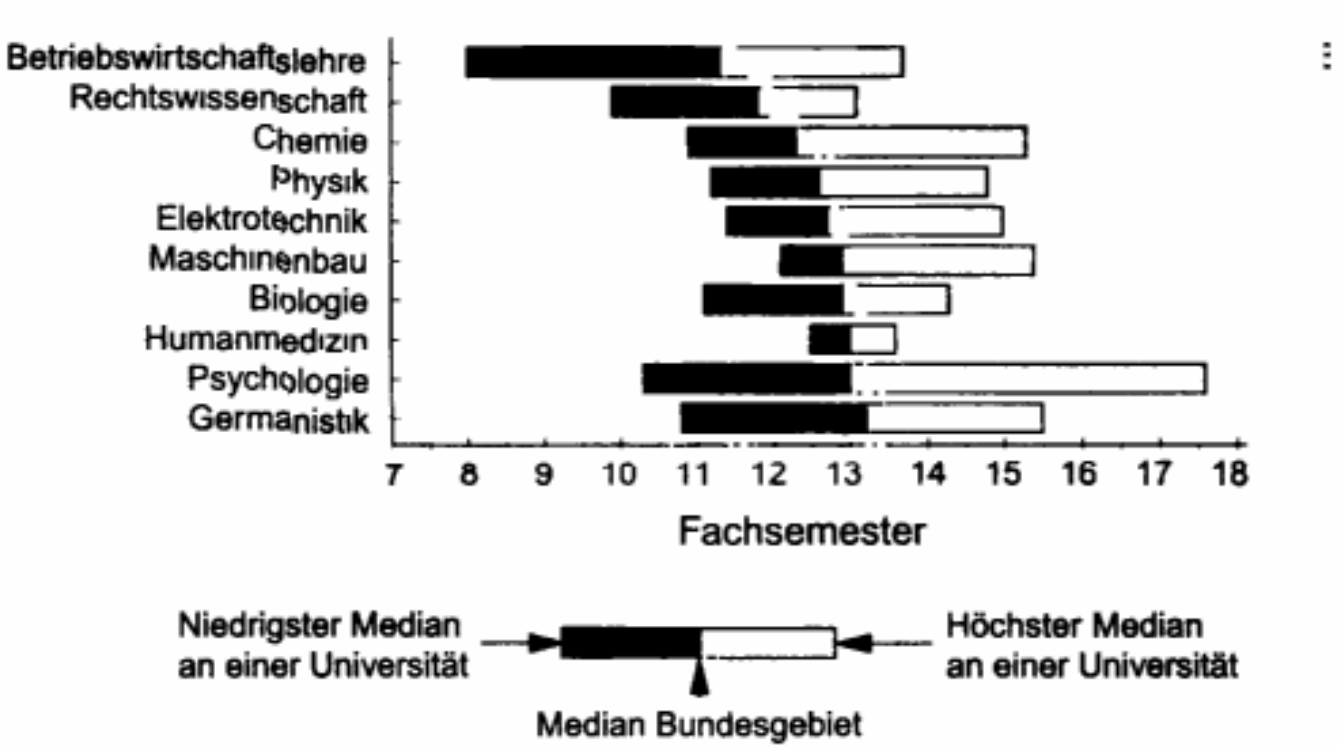


Abb. 4: *Durchschnittliche Fachstudiendauer von Absolventen verschiedener Universitäten aus den alten Bundesländern im Jahre 1991.*[9]

Es konnte jedoch gezeigt werden, dass sich individuelle Faktoren, wie ein hohes Alter bei der ersten Immatrikulation, eine berufliche Vorbildung und das Erachten einer kurzen Studienzeit als ein wichtiges Einstellungskriterium, studienzeitverkürzend auswirken. Wie diese Studie hat auch die Studie zu „Individuellen Determinanten von Studiendauer" von H. Giessen und A. Gold (1996), ergeben, dass der schulische Notendurchschnitt zwar die Studiendauer geringfügig beeinflusst, aber nicht in direktem Zusammenhang gebracht werden kann.

Lediglich die Wahl des Studienfaches wird durch die erbrachten Noten beeinflusst und das bei naturwissenschaftlichen, mathematischen und technischen Fächern stärker als in den Rechts- und Wirtschaftswissenschaften und dort wieder mehr als in den Geisteswissenschaften. Gute Schulnoten führen in den erstgenannten Fächern häufiger zu einer normalen Studierzeit.

Wie bereits erwähnt kann die Ausbildungsbiographie eines Studierenden ebenfalls einen positiven Einfluss auf die Studiendauer haben. Denn Studenten, die vor der Erlangung der Hochschulreife eine Berufsausbildung absolviert haben und anschließend erst an die Universität kamen, haben eine wesentlich kürzere Studienzeit. Die soziale Herkunft wirkt

[9] Giessen, H. und Gold, A. (1996) Individuellen Determinanten von Studiendauer. Lehr- und Lernprobleme im Studium.

sich ebenfalls, aber auf verschiedene Weise, auf die Länge des Studiums aus. Eine hohe soziale Herkunft wirkt sich an der Universität positiv aus, an der Fachhochschule eher negativ, während es sich mit niedriger sozialer Herkunft umgekehrt verhält. Solche Studenten haben an der Fachhochschule weniger Probleme als an der Universität und studieren deshalb dort schneller. Negative Auswirkungen auf die Studiendauer haben das Vorhandensein von Kindern (bei Frauen) und die Notwendigkeit einer Erwerbstätigkeit zur Finanzierung des Studiums und Lebensunterhalts. Kinderlose Frauen studieren im Schnitt ein Semester kürzer als welche mit Kind und eine Erwerbstätigkeit kann das Studium erheblich verlängern. Diese Faktoren zusammen erklären aber lediglich 25% der Varianz. Einer „informellen Studienunterbrechung" dagegen wird der größte Anteil zur Verlängerung eines Studiums – nämlich um gute zwei Semester – zugesprochen.

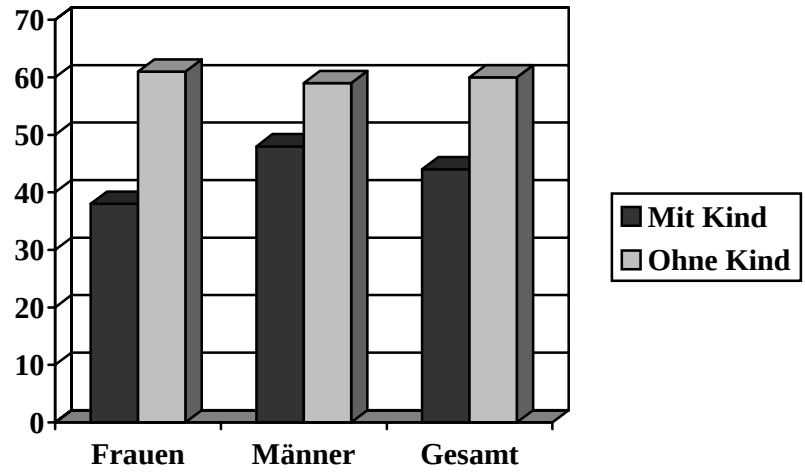


Abb. 5: Relative Fachstudiendauer von Hochschulabsolventinnen und -absolventen mit und ohne Kinder nach Geschlecht (Anteil der Befragten mit unterdurchschnittlich langer Fachstudiendauer; in %)[10]

Die institutionellen Faktoren beinhalten Gegebenheiten an der Universität, die für die Dauer eines Studiums verantwortlich sind, ohne dass die Studenten einen direkten Einfluss darauf haben. Dabei spielen die Studienorganisation, -struktur und -planung die wichtigsten Rollen, denn an diese müssen sich die Studenten unumgänglich halten. Diese drei erklären 74 % der Varianz in der Studiendauer, die auf institutionelle Faktoren zurückgeführt werden kann. Es hängt davon ab, ob die Veranstaltungen aufeinander aufbauen, die Regelung der Reihenfolge der Prüfungen und die Wiederholbarkeit der Prüfungen gut durchdacht sind. Eine zu hohe Regelungsdichte kann einerseits einschränkend wirken, wenn man studiert,

[10] Schaeper, H. und Minks, K.-H- (1997). Studiendauer – eine empirische Analyse ihrer Determinanten und Auswirkungen auf den Berufseintritt.

um sich zu bilden. Andererseits kann sie zu kürzeren Studienzeiten führen, wenn man zum Zwecke der Berufsausübung studiert. Außerdem gibt eine hohe Regelungsdichte eine Orientierungshilfe für das Studium. Ebenso ist ein guter sozialer Kontakt, sowohl zu den Lehrenden als auch zu anderen Studierenden, und klare Prüfungsanforderungen wichtig für ein schnelles Vorankommen im Studium. Schlechten Einfluss haben dagegen ein schlechtes Literaturangebot an der Universität, große Studentenzahlen in den Veranstaltungen und ein großer Hochschulort.

Ergänzend zu diesen Ergebnissen bringt eine Längsschnittstudie von H. Giessen und A. Gold weitere individuelle Determinanten der Studiendauer zutage. So wurde dort gezeigt, dass die Abiturnote unerheblich für die Studiendauer ist, jedoch für die Wahl des Studienfachs von großer Bedeutung, vorwiegend in den mathematischen, naturwissenschaftlichen und technischen Studienfächern. Auch zeigte sich in diesen Fachbereichen ein hoher Zusammenhang zwischen Studienzufriedenheit, -motivation und -bedingungen und der Studiendauer; mit zunehmender Studiendauer sinkt die Motivation und Zufriedenheit und die Bedingungen werden immer schlechter bewertet. Ein weiterer wichtiger Punkt, der die Studiendauer erheblich beeinflusst, zeigt auch wieder diese Studie, ist die Notwendigkeit einer Erwerbstätigkeit neben dem Studium. Hierbei wurden die erste und die zweite Studienhälfte, deren jeweilige Finanzierung und die Folgen, getrennt untersucht. Es wurde varianzanalytisch mit einem Chi-Quadrat-Test festgestellt, dass die zweite Studienhälfte eher durch eigene Mittel finanziert werden muss als die erste (da z.B. die staatliche Finanzierung abgelaufen ist) und daraus längere Studienzeiten und schlechtere Examensnoten resultieren.

Zur Herausstellung des tatsächlichen „Einflusses der Erwerbstätigkeit auf den Studienerfolg“ wurde wiederum von H. Bardstätter und A. Farthofer (2003) eine gesonderte Studie durchgeführt. Dabei stellte sich nach regressionsanalytischer Auswertung der Ergebnisse des Fragebogens heraus, dass ein Zeitaufwand von 19 Stunden und mehr pro Woche zu erheblichen Beeinträchtigungen des Studienerfolgs führt. Der negative Einfluss einer Erwerbstätigkeit lässt die Anzahl der abgelegten Prüfungen pro Semester, den Notendurchschnitt, die Studienzufriedenheit und die Stabilität der Studienwahl sinken und damit das Abbruchrisiko steigen. Dabei lassen sich keine Zusammenhänge zwischen dem Alter oder der Art (fachfremd vs. fachähnlich) der Erwerbstätigkeit erkennen. Man erhält jedoch das Resultat, dass die für das Studium aufgewendete Zeit einen geringeren positiven Einfluss auf den Studienerfolg hat, als die für die Erwerbstätigkeit aufgewendete einen negativen. Beide Variablen zusammen erklären 35 % der Varianz des Notendurchschnitts.

Gründe für eine Erwerbstätigkeit sind weniger die sozio-ökonomischen Verhältnisse aus denen die Studierenden kommen, sondern vielmehr Persönlichkeitseigenschaften. Es hat sich gezeigt, dass Extraversion gepaart mit dem Wunsch nach Unabhängigkeit signifikant zu einem höheren Zeitaufwand für die Erwerbstätigkeit führt.

Studiendauer und Berufseintritt

Während Anfang der 80er Jahre der Studiendauer als Einstellungskriterium noch eine untergeordnete Rolle zugeschrieben wurde, hat sich deren wahrgenommene und tatsächliche Wichtigkeit wesentlich erhöht. Die Bedeutung eines kurzen Studiums wird nach wie vor in

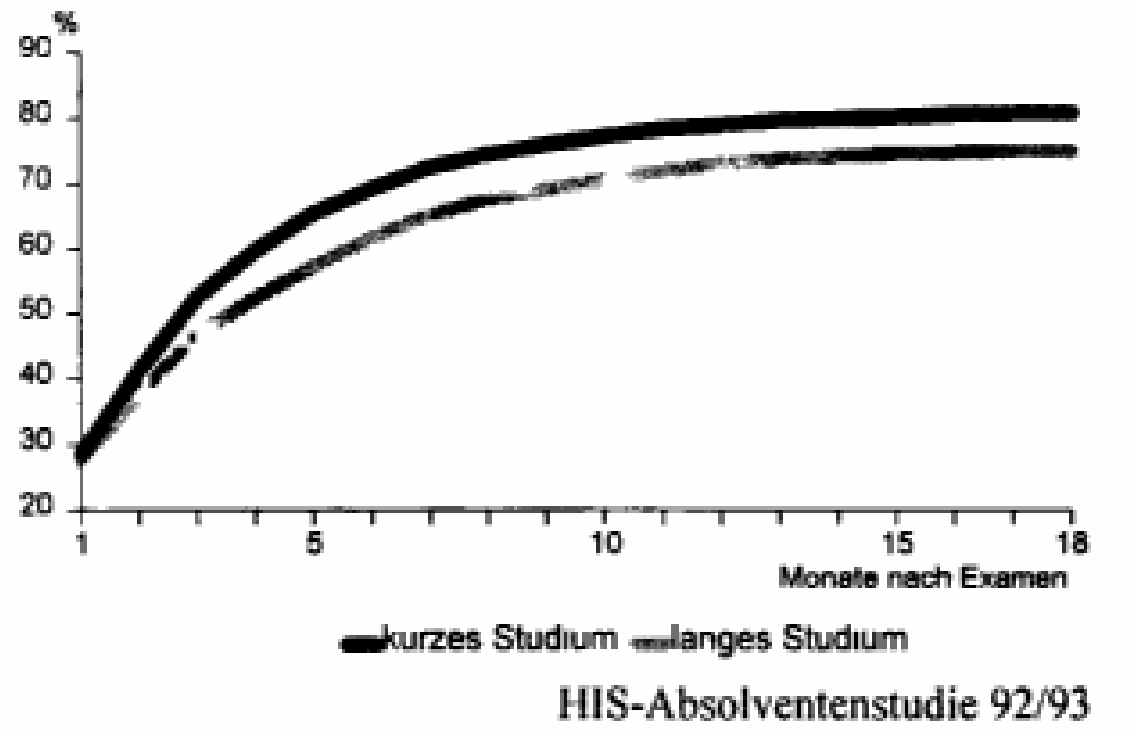


Abb. 6: Aufnahme der ersten regulären Erwerbstätigkeit nach relativer Fachstudiendauer (Anteil der Befragten, die innerhalb der ersten Monate nach Examen eine reguläre Erwerbstätigkeit aufgenommen haben; in %, kumuliert)[11]

Studiengängen, die in einen Beruf in der privaten Wirtschaft münden, höher eingeschätzt als in Berufen, die in den öffentlichen Dienst führen. Die folgenden Analysen von H. Schaeper und K.-H. Minks (1997) wurden mithilfe eines Fragebogens durchgeführt, wobei die Befragten Angaben zu ihrer Situation im Zeitraum bis 18 Monate nach dem Abschluss machen sollten. In dem gleichen Maße, in dem die Wichtigkeit der Studiendauer wahrgenommen wird, kann auch der Eintritt in das Berufsleben wahrgenommen werden.

[11] Schaeper, H. und Minks, K.-H- (1997). Studiendauer – eine empirische Analyse ihrer Determinanten und Auswirkungen auf den Berufseintritt.

Absolventen, die schnell studiert haben, finden in der Regel auch schneller und häufiger einen Beruf nach ihrem Examen als langsame Studenten.

Es ist auch wieder festzustellen, dass dies eher für die natur- und wirtschaftswissenschaftlichen Absolventen gilt als für die Geisteswissenschaften und Lehramtsfächer. Auch ist die Stellensuche der guten Absolventen effizienter und erfolgreicher als die der schlechten, wenn man die Häufigkeit von Bewerbungsgesprächen und Anzahl der versendeten und anschließend positiv beantworteten Bewerbungen als Indikatoren nimmt. Eine Ausnahme stellen hier die universitären Abschlüsse in BWL und Biologie dar. Erstaunlicherweise spielt die Note des Abschlusszeugnisses nur eine untergeordnete Rolle für die Stellensuche. Absolventen mit guten Noten haben nur geringfügig früher eine Erwerbstätigkeit aufgenommen als solche mit einer schlechten Note. Mehr als Studiendauer und Examensnote zählt jedoch das Vorhandensein einer abgeschlossenen Berufsausbildung vor dem Studium für einen schnellen Eintritt in eine reguläre Erwerbstätigkeit. Außerdem haben Absolventen der meisten Fachbereiche, die schnell studiert haben, bessere Positionen, ein höheres Einkommen, stabilere Beschäftigungsverhältnisse und sind mit ihrer beruflichen Situation zufriedener.

Fazit

Zur Auswahl geeigneter Studenten lassen sich gut die Schulnoten heranziehen. Um einen weiteren Studienerfolg vorsagen zu können, eigenen sich die Noten des ersten und zweiten Studiensemesters sehr gut als Prädiktoren. Neben diesen Faktoren spielt selbstverständlich ebenfalls die Motivation eine große Rolle für den Erfolg in einem Studium. Ein Test mit fachspezifischen Aufgaben kann nicht den Studienerfolg vorhersagen. Er kann aber im Rahmen der Studienberatung vor Antritt des Studiums eine gute Orientierung sein, welche Aufgaben während des Studiums auf einen zukommen und ob man in der Lage und willig ist sich ihnen zu stellen. Damit kann einem späteren Studienabbruch vorgebeugt werden. Der Studienabbruch an sich ist schwer vorherzusagen, da er auf vielen meist persönlichen Faktoren beruht.

Den Studienerfolg über die Studiendauer vorherzusagen, gestaltet sich ebenfalls nicht einfach. Es gibt viele Faktoren, die die Studiendauer beeinflussen. Auf die individuellen Faktoren hat die Universität nahe zu keinen Einfluss. Hier liegt es beim Studenten diese Faktoren zu erkennen und wenn möglich und nötig zu beheben. So wirken sich eine Erwerbstätigkeit neben dem Studium und das Vorhandensein eines Kindes meist negativ auf die Studierzeit aus. So sollte die aufgewendete Zeit für die Erwerbstätigkeit 19 Stunden pro Woche nicht übersteigen, da sonst das Studium erheblich unter diesem Zeitverlust für studiumsorientierte Aufgaben leidet. Wenn jedoch der Student aufgrund seiner Einstellung die Länge eines Studiums nicht als wichtig erachtet, sind die Einflussmöglichkeiten der Universität begrenzet. Es hat sich nicht gezeigt, dass ein Langes Studium per se ein Nachteil für einen schnellen Berufseintritt ist. Vor allem in den geisteswissenschaftlichen

Bereichen – so auch in der Psychologie – werden Absolventen mit einem langen Studium bevorzugt oder zumindest nicht benachteiligt. Die Abschlussnoten sollte sich jedoch mindestens im durchschnittlichen Bereich befinden.

Auch die Universitäten können dazu beitragen, dass ein Studium nicht länger dauert als nötig. Dafür sollte der Verlauf des Studiums gut durchdacht sein. Die Organisation und Struktur des Studienganges sollte durchsichtig und klar und die Planung einfach sein. Eine hohe Regelungsdichte mag die Entfaltungsmöglichkeiten einschränken, hat aber den Vorteil, dass die Orientierung leicht fällt und Weg durch das Studium vorgegeben ist. Eine geringe Zahl von aufeinander aufbauenden Veranstaltungen sowie Studenten in Veranstaltungen begünstigen ein zügiges Fortschreiten im Studium genauso wie die Wiederholbarkeit von Prüfungen während des Semesters. Eine gute Beratung sowie ein guter sozialer Kontakt sowohl zu den Lehrenden als auch anderen Studenten, sind ebenfalls wichtig für ein zügiges Vorankommen im Studium. Weiter Voraussetzungen für einen guten und schnellen Abschluss sind eine gute Ausstattung von Lehrmitteln und ein gutes als auch reiches Literaturangebot.

Literaturverzeichnis

Brandstätter, H. & Farthofer (2003a). Einfluss von Erwerbstätigkeit auf den Studienerfolg. *Zeitschrift für Arbeits- und Organisationspsychologie, 47,* 134-145.

Brandstätter, H. & Farthofer, A. (2003b). Erste Prüfungen - weiterer Studienerfolg. Psychologie in Erziehung und Unterricht, 50, 58-70.

Giesen, H. & Gold, A. (1996). Individuelle Determinanten der Studiendauer. In J. Lompscher & H. Mandl (Hrsg.), *Lehr- und Lernprobleme im Studium: Bedingungen und Veränderungsmöglichkeiten*. Bern: Huber.

Schaeper, H. und Minks, K.-H. (1997). *Studiendauer - eine empirische Analyse ihrer Determinanten und Auswirkungen auf den Berufseintritt. HIS-Kurzinformationen, A1/97.* Hannover.

Kompetenzmodelle am Beispiel der dritten internationalen Mathematik- und Naturwissenschaftsstudie TIMSS III

Eva Riedmüller

Einleitung

Die Entwicklung von Kompetenzstufenmodellen für bestimmte Aufgabenbereiche wie z.B. die Mathematik, kann die Zuordnung von Personen zu verschiedenen Fähigkeitsniveaus ermöglichen. Durch Testverfahren, die die Kompetenzen mithilfe von Aufgaben ermitteln, lassen sich Aussagen machen über die durchschnittliche Fähigkeit einer Jahrgangsstufe, einer Klasse oder auch einer einzelnen Person.

Verfahren zur Studierendenauswahl sind noch nicht entwickelt, die vorhandenen Tests der Fähigkeitsstudien wie z.B. PISA könnten hierzu jedoch als Anregung dienen.

Im folgenden Text werden die Hintergründe zur allgemeinen Kompetenzstufenmodell- Idee erläutert, sowie die Entwicklung eines solchen Modells am Beispiel der third international mathmatic and science study TIMSS III dargestellt.

Kompetenzen sind latente personelle Merkmale, „deren Ausprägung sich in unterschiedlichen Situationen bei unterschiedlichen Anforderungen zeigt“ (Klieme, Baumert, Köller & Bos, 2000).

Die Einteilung eines Kompetenzkontinuums in einzelne Fähigkeitsstufen basiert auf der Annahme, dass Personen, die Aufgaben eines höheren Schwierigkeitsniveaus lösen können, auch alle Aufgaben von geringerer Schwierigkeit meistern. Diese Annahme ist empirisch bestätigt. Die dritte internationale Mathematik- und Naturwissenschaftsstudie TIMSS III untersuchte die Fähigkeiten von Schülen am Ende der Pflichtschulzeit und am Ende der gymnasialen Oberstufe im internationalen Vergleich, ähnlich der PISA Studie.

Im Folgenden wird zunächst auf das Konzept der Kompetenz eingegangen. Weiterhin werden Vorstellungen darüber beschrieben, welche Kompetenzen im Unterricht erworben werden sollten.

Das Kompetenzkonzept

Bildungsstandards stellen genau wie Lehrpläne einen Kompromiss zwischen der Orientierung an fachlicher Systematik, an funktionalen Anforderungen der Lebens- und Arbeitswelt und an den Lernvoraussetzungen und Entwicklungsbedürfnissen der Lernenden

dar. In der klassischen bildungstheoretischen Diskussion (insbesondere im Lehrplan Humboldtscher Prägung) wurden diese drei Bereiche einheitlich behandelt, Bildung verstand man als „die Aneignung unterschiedlicher Zugänge zur Welt“, die in den verschiedenen Schulfächern vermittelt werden sollte. Ab den 70er Jahren formulierte man detaillierte fachbezogene Lernziele. Mit der Expansion des Bildungswesens ließen sich Bildungsziele allerdings nicht mehr rein fachspezifisch definieren. Besonders in der beruflichen Bildung zeigte sich, dass Anforderungen nicht ausschließlich inhaltlicher Natur waren. In den 1980er und 90er Jahren vollzog sich ein Wandel: Sowohl in berufsbezogenen als auch in schulischen Lehrplänen wurden immer mehr grundlegende Zieldimensionen beschrieben wie die

- Fähigkeit zum kritischen Denken,
- Problemlösefähigkeit und
- Kooperationsfähigkeit.

In dieser Zeit entwickelte sich eine Tendenz zur Idee von transferierbaren Schlüsselqualifikationen. Zeitzeugen dieser Entwicklung sind Konzepte wie „Lernen lernen“, „eigenverantwortliches Arbeiten„ oder „der geheime Lehrplan“.

Die Literacy Diskussion

Vergleichbar zum Konzept der Schlüsselqualifikationen ist im angelsächsischen Bereich die Literacy Diskussion. Neben dem Erlernen der Muttersprache in Wort und Schrift (Literacy = Fähigkeit zu lesen und zu schreiben) als erste Basiskompetenz stellt auch die Aneignung von Mathematik als formalisierte Sprache inzwischen ein selbstverständliches Kommunikationsmittel in einigen Berufen und im wissenschaftlichen Bereich dar. Beide Sprachen sind kulturelle Werkzeuge. Um der mathematischen Basisqualifikation den gleichen Rang zukommen zu lassen, spricht man metaphorisch von mathematischer Literalität oder Mathematics Literacy.

Der National Council of Teachers of Mathematics (NCTM) formulierte fünf allgemeine Zieldimensionen mathematischer Literalität, die durch den Mathematik Unterricht vermittelt werden sollten:

1. Wertschätzung von Mathematik (to value mathematics)
2. Vertrauen in die eigene Fähigkeit, mit Mathematik umgehen zu können (selfconfidence to do mathematics)
3. Anwendung mathematischer Kenntnisse auf innermathematischer und außermathematischer Aufgabenstellung (to solve mathematical problems)
4. Kommunikation mithilfe der Mathematik (to communicate mathematically)
5. mathematisches Denken (to reason mathematically)

(NCTM, 1989 in Klieme, 2000).

Unter den Begriff der Mathematic Literacy fallen demnach neben mathematischen Kompetenzen und Haltungen auch soziale Fähigkeiten.

Die Anforderungen an den Mathematikunterricht werden im Folgenden konkretisiert. Der NCTM definiert genau, zu welchem Qualifikationserwerb dem Schüler verholfen werden soll:

- Vorbereitung auf offene Aufgabenstellungen (...)
- Fähigkeit, die Anwendbarkeit mathematischer Konzepte und Modelle auf alltägliche und komplexe Problemstellung zu erkennen,
- Fähigkeit, die einem Problem zu Grunde liegende mathematische Struktur zu sehen,
- Fähigkeit, Aufgabenstellungen in geeignete Operationen zu übersetzen, und
- ausreichende Kenntnis und Beherrschung von Lösungsroutinen."

(NCTM, 1989 in Klieme, 2000)

Ähnliche Ziele für den naturwissenschaftlichen Unterricht formulierte die American Association for the Advancement of Science, die Benchmarks for Science Literacy (AAAS):

- Vertrautheit mit der natürlichen Welt und Kenntnis ihrer Verschiedenheit und Einheit
- Verständnis zentraler naturwissenschaftlicher Konzepte und Prinzipien,
- Kenntnis der Interdependenz von Naturwissenschaften, Mathematik und Technik,
- epistemologische Vorstellungen von der konstruktiven Natur der Mathematik, den Naturwissenschaften und der Technologie sowie Kenntnis ihrer Stärken und Grenzen,
- Verständnis der Grundzüge naturwissenschaftlichen Denkens,
- Anwendung von naturwissenschaftlichem Wessen auf Sachverhalte des individuellen und sozialen Lebens.

(AAAS, 1994 in Klieme et al., 2000).

Wie am Beispiel von NCTM und AAAS und zu sehen ist, werden hier, anstatt einer Auflistung formaler, inhaltlicher Bildungsziele, funktionale Zielvorstellungen formuliert. Der Grundgedanke ist in beiden Fällen, der Tiefe des Verständnisses den Vorrang gegenüber der Breite des Stoffes zu geben. Weniger ist mehr.

Zur Ermittlung des Fähigkeitsniveaus einer Person

Um das Fähigkeitsniveaus einer Person zu ermitteln, bedarf es Tests, die sich im Falle der TIMSS III aus Mathematik- bzw. Naturwissenschaftsaufgaben zusammensetzen.

Die Aufgaben gliedern sich in Gruppen unterschiedlicher Schwierigkeit, so dass eine Trennung von guten und weniger guten Bearbeitern bzw. eine Zuordnung der Personen auf bestimmte Kompetenzstufen möglich wird.

Um geeignete Aufgaben für die verschiedenen Schwierigkeitsgruppen zu entwickeln, bedarf es einiger Vorüberlegungen.

Zunächst stellt sich die Frage, welche Faktoren die Lösungswahrscheinlichkeit für eine Aufgabe beeinflussen. Dem Rasch-Modell zufolge ergibt sich diese aus Itemparametern und Personenparametern. Unter Itemparametern versteht man die Schwierigkeit einer Aufgabe, die Personenparameter sind die Fähigkeiten des Bearbeiters. Auf beide Parameter wird im Folgenden näher eingegangen.

Itemparameter

- Nach Williams und Clarke (1997) wird die Komplexität einer mathematischen Aufgabe durch sprachliche Merkmale,
- Kontextmerkmale,
- Art und Variabilität der Darstellungsformen (Graphen, Formeln usw.) ,
- Art und Anzahl der erforderlichen mathematischen Operationen,
- Art und Verknüpfung von mathematischen Begriffen und durch die
- „intellektuelle Komplexität“ der Aufgabe bedingt.

Verschiedene Studien versuchten bis Ende der 70er Jahre, Aufgabenschwierigkeit durch inhaltliche und sprachliche Merkmale zu erklären, auch in der Art des erwarteten Lösungsprozesses vermutete man einen die Schwierigkeit bestimmenden Faktor.

Barnett (1979) belegte diese Vermutung in einer Studie. Die Art und Reihenfolge der erforderlichen mathematischen Operationen klärten den größten Teil der Schwierigkeitskennwerte mathematischer Aufgaben auf, sprachliche Merkmale zeigten hier unterschiedliche Effekte darauf haben.

Aufgabenmerkmale wie sprachliche Ausführung, Darstellungsform und Anzahl und Art der erforderten Operationen sind objektiv, sie unterscheiden sich nicht bei verschiedenen Bearbeitern.

Klieme (1989) fand in einer Untersuchung der Schwierigkeit mathematischer Sachaufgaben im Test für medizinische Studiengänge heraus, dass nur ein Drittel der Schwierigkeitsvarianz durch objektive Aufgabenmerkmale erklärt wird.

Die Bearbeitungsprozesse scheinen demnach eine große Rolle für die Schwierigkeit einer Aufgabe zu spielen, die im nächsten Abschnitt behandelt werden.

Personenparameter

Seit den 70ger Jahren gewannen kognitionspsychologische Prozessanalysen bei der Untersuchung mathematisch - naturwissenschaftlicher Aufgaben an Bedeutung.

In mehreren Studien wurde nachgewiesen, dass die Art der Aufgabenbearbeitung von verschiedenen Faktoren abhängt:

- Art und Struktur des individuellen Vorwissens,
- die Einstellungen und Erwartungen des Bearbeiters sowie die
- sozialen Kontextbedingungen.

Die Art des Vorwissens wirkt sich darin aus, das Experten qualitativ anders arbeiten als Anfänger. Sie setzen Lösungsschemata zur Bearbeitung ein (Top-Down- Strategien) während Anfänger die Aufgabe vom Grund her angehen (Bottom-up) und sehr viel komplexere Lösungsstrategien anwenden (Reimann & Chi in Gilhooly et al., 1989). Im Gegensatz zu früheren Modellen gehen Lesgold, Lajoie, Logan, Eggan (in Frederiksen, Glaser, Lesgold, Shafto, 1990) von einer Auswirkung der vom jeweiligen Vorwissen abhängigen Bearbeitungsstrategien auf die Aufgabenlösung ausgegangen. Es sind demnach nicht ausschließlich die objektiven Merkmale einer Aufgabe, die deren Schwierigkeit für den Bearbeiter bestimmen.

Zur Einordnung von Aufgaben anhand ihrer Schwierigkeit gibt es verschiedene Ansätze, auf die nun kurz eingegangen wird.

Stein, Grover und Henningsen (1996) benennen drei Aufgabenmerkmale. Das erste Merkmal ist, ob Formeln, Algorithmen oder Prozeduren mit oder ohne Verbindung zu konzeptuellem Wissen angewendet werden können. Das zweite Aufgabenmerkmal nennen sie die „Offenheit der Lösungsweg“. Hierbei ist die Frage, ob unterschiedliche Strategien zur Lösungsfindung angewendet werden können, und ob eine Erklärung des Ergebnisses verlangt wird. Die dritte und höchste Anforderungsstufe nennen sie „doing mathmatics“, mathematisches Schlussfolgern, Argumentieren, Problemlösen und Entdecken von Mustern.

Problemlösendes Denken erfordert den Aufbau eines mentalen Situationsmodells, d.h. das Verstehen der Problemsituation. Art, Umfang und Vernetztheit des Situationsmodells sind wesentliche Komponenten der Aufgabenkomplexität. Wie leicht das Situationsmodell erstellt werden kann hängt unter anderem von den Darstellungsformen einer Aufgabe ab, z.B. als Text, in Diagrammen etc. Diese Repräsentationsformen beeinflussen die Schwierigkeit der Aufgabenbearbeitung (Stern, 1998).

Für naturwissenschaftliches, insbesondere physikalisches Denken besteht ein weiteres Anforderungsmerkmal, das Überwinden von „Fehlvorstellungen“. Schüler haben diese Fehlvorstellungen über Physik aufgrund von Alltagserfahrungen. Zuverlässige Messverfahren zu diesem Faktor stehen noch aus.

Wie in diesem Abschnitt gezeigt wurde, bestimmen einige Faktoren den Lösungsprozess und damit den Bearbeitungserfolg.

Die Entwicklung eines Kompetenzstufenmodells am Beispiel der TIMSS III

Die Schwierigkeit einer Aufgabe hängt, wie schon erwähnt, sowohl von den objektiven Merkmalen einer Aufgabe als auch von deren Anforderungen an die Fähigkeiten einer Person und der individueller Ausprägung dieser Fähigkeiten ab.

Tests mit monotonen Item- Charakteristik- Funktionen

Die Untertests der TIMSS III gehören dieser Testkategorie an. Bei Tests mit monotonen Item- Charakteristik- Funktionen steigt die Lösungswahrscheinlichkeit eines Items mit der latenten Fähigkeit (Kompetenz) des Bearbeiters. Ein erfolgreicher Testbearbeiter unterscheidet sich von einem weniger erfolgreichen dadurch, das er auch anspruchsvollere Aufgaben lösen kann.

Die Testaufgaben dienen, neben der Ermittlung der Fähigkeiten einer Person für einen Aufgabentyp, auch dazu, einen Rückschluss auf universell definierte Fähigkeiten, die Mathematics bzw. Science Literacy zu ermöglichen.

Voraussetzungen zur Messung von Kompetenz

Zur Anwendung von Indikatoren für Kompetenzen müssen verschiedene Vorraussetzungen erfüllt sein:

1. Die Skalen müssen eine ausreichende Breite von Anforderungen, Inhalten und Aufgabenformaten berücksichtigen.
2. Es muss durch testtheoretische Analysen belegt sein, dass homogene Skalen gebildet werden können.

3. Die postulierte Fähigkeitsdimension muss als wissenschaftliches Konstrukt definiert und beschrieben, und in Modellvorstellungen eingebettet sein.

Die TIMSS erfüllt diese Bedingungen, die Rasch- Skalierung der Skalen gewährleistet die zweite Vorraussetzung. Informationen zur Rasch- Skalierung finden sich bei Steyer und Eid (2001).

Für jedes Item werden Schwierigkeitsparameter und die Ausprägung verschiedener Anforderungsmerkmale ermittelt. Die Variation der Schwierigkeiten lässt sich über Regressions- bzw. Varianzanalysen auf die Anforderungsmerkmale zurückführen.

Die Anforderungsmerkmale der TIMSS III- Items

Inhaltlichen Anforderungen

Die inhaltlichen Anforderungsmerkmale der TIMSS III- Aufgaben wurden von den Testautoren definiert.

Sie unterscheiden Aufgaben nach

- verschiedene Anforderungskategorien,
- Inhaltsbereichen, die jeweils ein Stoffgebiet erfassen und
- dem Format der verlangten Antwort. Diese differenzieren sie in multiple choice Format, Kurzantworten wie etwa eine Zahl oder ein Stichwort, und die erweiterte Anwort, in der z.B. ein Beweis durchgeführt oder eine Versuchsanordnung dargestellt werden soll.

Für den Grundbildungsteil wurden drei weitere Anforderungsmerkmale hinzugefügt.

- Offenheit der Aufgabenstellung
- Komplexität des Lösungsprozesses und das
- Niveau der Grundbildung nach dem Literacy- Konzept von Shamos (1995) und Bybee (1997).

Anforderungen an die Fähigkeit des Bearbeiters

Zur Ermittlung der Fähigkeitsanforderungen wurden die TIMSS III- Items 10 Mathematik- und 9 Physikexperten vorgelegt, die diese aus Sicht des typischen Bearbeiters der Oberstufe (Grund- bzw. Leistungskurs) betrachten und dann beurteilen sollten, wie wichtig folgende Merkmale für die Bearbeitung der Aufgabe sind.

Gemeinsame Merkmale:

G1: Kenntnis von Definitionen, mathematischen Sätzen bzw. physikalischen Gesetzen

G2: Qualitatives Verständnis mathematischer Bzw. physikalischer Begriffe.

G3: Rechnen (Arithmetisches Operieren)

G5: Interpretation von Diagrammen (Koordinatendarstellung oder statistische Diagramme)

G6: Textverständnis

G7: Bildliches Vorstellen

G8: Problemlöseprozesse

Ergänzend für die Mathematikaufgaben:

M1: Formalisierung

M2: Interpretation von Anwendungssituationen

M3: Umstrukturierung / Flexibilität

M4: Prinzipien mathematischen Denkens

Ergänzend für die Physikaufgaben:

P1: Verständnis formalisierter Gesetze

P2: Verständnis für funktionale Zusammenhänge

P3: Verständnis von Alltagssituationen

P4: Verständnis für experimentelle Situationen

P5: Verständnis symbolischer Zeichnungen

P6: Überwindung von Fehlvorstellungen

Diese Merkmale, über die ein Bearbeiter verfügen kann oder auch nicht, wurden von den Experten mit den Kodierungen „0= keine Bedeutung“, „1= untergeordnete Bedeutung“, „2= hohe Bedeutung“ oder „3= entscheidend für Erfolg oder Misserfolg“ versehen. Die Kodierung erfolgte, für wie wichtig sie die Ausprägung des jeweiligen Merkmals für die Lösung der einzelnen Aufgaben erachteten.

Unter Berücksichtigung der inhaltlichen Anforderungen (Itemparameter) und der zur Lösung einer Aufgabe benötigten Fähigkeiten des Bearbeiters (Personenparameter) ließen sich nun die Schwierigkeiten der Aufgaben berechnen. Zum Vorgehen dieser Methode der Item- Response Theorie bzw. Probabilistischen Testtheorie sei an dieser Stelle auf Moosbrugger und Müller (1997) verwiesen.

Die Kompetenzstufen

Die Autoren der TIMSS III haben für alle vier Untertest der mathematischen und naturwissenschaftlichen Aufgaben für das Ende der Pflichtschulzeit bzw. das Ende der gymnasialen Oberstufe Kompetenzstufen formuliert, die durch die Fähigkeitskennwerte (Scores) definiert sind. Die Kennwerte werden auf einer Skala mit einem Mittelwert von 500 und einer Standardabweichung von 100 angegeben. Der Mittelwert und die darunter bzw. darüber liegenden Stufen (z.B. 400 und 600) wurden inhaltlich charakterisiert anhand der Leitfrage, welche Aufgaben von einer Person mit dem Score von beispielsweise 500 mit ausreichender Sicherheit gelöst werden, jedoch nicht von Personen, die einen niedrigeren Fähigkeitskennwert (z.B. 400) erreichen. Ausreichende Sicherheit bedeutet hier eine Chance von 2:1, oder mit anderen Worten eine Lösungswahrscheinlichkeit von mindestens 65%.

Im Folgenden werden die Kompetenzstufen für den Mathematik- und Naturwissenschaftsuntertest der TIMSS III, und im Anschluss diese der Untertests für die Oberstufen in denselben Fachgebieten vorgestellt.

Fähigkeitsniveaus für die Untertests am Ende der Pflichtschulzeit

Wie im vorhergehenden Abschnitt erwähnt wurden jeder Stufe Aufgaben zugeordnet, die von Personen auf dieser Stufe mit hinreichender Sicherheit (65 Prozent) gelöst werden können, jedoch nicht auf der darunter liegenden Stufe. Dies ermöglicht eine trennscharfe Charakterisierung der angrenzenden Kompetenzstufen.

Die Kompetenzstufen der Grundbildungstest werden in Tabelle 1 zusammengefasst.

Tabelle 1: Kompetenzstufen für die TIMSS III – Grundbildungstests (in Anlehnung an Klieme et al. (2000))

Gruppe (Score)	Fähigkeitsniveau (Kompetenzstufe)	Inhaltliche Charakterisierung der jeweils erfolgreich lösbaren Aufgaben	
		Mathematik	**Naturwissenschaften**
< 400	Maximal Stufe 1	Alltagsbezogene Schlussfolgerungen	Naturwissenschaftliches Alltagswissen
401 – 500	Maximal Stufe 2	Anwendung von einfachen Routinen	Erklärung einfacher alltagsnaher Phänomene
501 – 600	Maximal Stufe 3	Bildung von Modellen und Verknüpfungen von Modellvorstellungen	Anwendung elementarer Operationen naturwissenschaftlicher
> 600 oder höher	bis zu Stufe 4	Mathematisches Argumentieren (insbesondere anhand graphischer Darstellungen)	Verfügung über grundlegende naturwissenschaftliche Fachkenntnisse

Die Stufen werden folgendermaßen definiert:

Kompetenzstufen für Mathematik im Grundbildungstest

- *Stufe 1 (Score 400): Alltagsbezogenes Schlussfolgern*

  Auf dieser Stufe werden keine expliziten mathematischen Operationen verlangt, weder Rechnungen noch Formalisierungen. Einfache Überlegungen wie z.B. „Je mehr Schritte jemand braucht, um eine bestimmte Entfernung zu überwinden, desto kleiner ist seine Schrittlänge“ (Klieme et al., 2000), können auf Stufe 1 vollzogen werden.

- *Stufe 2 (Score 500): Anwendung von einfachen Routinen*

Die Aufgaben dieser Stufe verlangen einfache Proportionalitätsüberlegungen, Prozentrechnung oder Flächenberechnungen.

- *Stufe 3 (Score 600): mathematisches Modellieren auf einfachem Niveau*

Unterschiedliche Operationen wie z.B. Volumenbestimmung und Verhältnisrechnung, die nicht in der Aufgabe angeführt sondern selbst zu erschließen sind, müssen hier miteinander verknüpft werden.

- *Stufe 4 (Score 700): mathematisches Argumentieren*

Hier wird z.B. das Erstellen und Interpretieren eines relativ komplexen Diagramms verlangt.

Kompetenzstufen für Naturwissenschaft im Grundbildungstest

- *Stufe 1 (Score 400): Naturwissenschaftliches Alltagswissen*

Vertraute Assoziationen aus Alltagserfahrungen genügen, um Aufgaben auf dieser untersten Stufe zu lösen wie z.B. dass der menschliche Körper an heißen Tagen Wasser verliert.

- *Stufe 2 (Score 500): Fähigkeit, alltagsnahe Phänomene in einfacher Weise zu erklären:*

Das Vorhandensein einfacher naturwissenschaftliche Modellvorstellungen ist hier hilfreich, wenn auch nicht zwingend erforderlich.

- *Stufe 3 (Score 600): Elementare naturwissenschaftliche Modellvorstellungen*

Beispielsweise das Verständnis der Vorgänge in einem Ökosystem oder das Konzept des Drucks in der Physik ist auf dieser Stufe vorhanden.

- *Stufe 4 (Score 700): Anwendung und Argumentation mit grundlegenden naturwissenschaftlichen Fachkenntnissen*

Begriffe wie beispielsweise „potentielle Energie“ oder „kinetische Energie“ können verstanden und in Diskussionen verwendet werden, das Erklären eines Stromkreises oder eines photosynthetischen Experimentes können Personen auf dieser höchsten Stufe leisten.

Kompetenzstufen für Mathematik am Ende der gymnasialen Oberstufe

Die Kompetenzstufen für den Bereich Mathematik werden in Tabelle 2 zusammengefasst.

Tabelle 2: Kompetenzstufen für den Testbereich Mathematik am Ende der Oberstufe (in Anlehnung an Klieme et al., 2000)

Gruppe (Score-Bereich)	Fähigkeitsniveau (Kompetenzstufe)	Inhaltliche Charakterisierung der jeweils erfolgreich lösbaren Aufgaben Mathematik
≤ 400	Maximal Stufe 1	elementares Schlussfolgern
401-500	Maximal Stufe 2	Anwendung einfacher mathematischer Begriffe und Regeln
501-600	Maximal Stufe 3	Anwendung von Lerninhalten der Oberstufe
> 600	bis zu Stufe 4 oder höher	selbständiges Lösen mathematischer Probleme auf Oberstufenniveau

Die Stufen werden wie folgt erklärt:

- *Stufe 1 (Score 400): Elementares Schlussfolgern*

  Elementares Schlussfolgern wird lediglich auf der Basisch einfachster arithmetischer Operationen verlangt, formale mathematische Argumentation werden nicht erwartet.

- *Stufe 2 (Score 500): Anwendung einfacher mathematischer Begriffe und Regeln, die kein Verständnis von Konzepten der Oberstufenmathematik voraussetzt*

  Auf Stufe 2 werden Fachbegriffe in der Aufgabenstellung verwendet, eine Lösung ist jedoch durch einfaches Rechnen und Schlussfolgern möglich.

- *Stufe 3 (Score 600): Anwendung von Lerninhalten der Oberstufe im Rahmen typischer Standardaufgaben*

  Zum Lösen der Aufgaben dieser Stufe sind Wissensinhalte der Sekundarstufe II notwendig.

- *Stufe 4 (Score 700): Selbständiges Lösen mathematischer Probleme auf Oberstufenniveau*

Hier wird neben Kenntnissen der Oberstufenmathematik vorwiegend eigenständiges Problemlösen und Argumentieren sowie auch die Kombination verschiedener Sichtweisen auf eine Aufgabe gefordert.

Kompetenzstufen für Naturwissenschaften am Ende der gymnasialen Oberstufe

Die Kompetenzstufen für den Bereich Physik werden in Tabelle 3 zusammengefasst.

Tabelle 3: Kompetenzstufen für den Testbereich Physik am Ende der Oberstufe (in Anlehnung an Klieme et al., 2000)

Gruppe (Score-Bereich)	Fähigkeitsniveau (Kompetenzstufe)	Inhaltliche Charakterisierung der jeweils erfolgreich lösbaren Aufgaben Physik
≤ 450	Maximal Stufe 1	Lösen von Routineaufgaben mit Mittelstufenwissen
451-550	Maximal Stufe 2	Anwendung von Faktenwissen
551-650	Maximal Stufe 3	Anwendung physikalischer Gesetze (Größengleichungen)
651-750	Maximal Stufe 4	selbständiges fachliches Argumentieren und Problemlösen
> 751	bis zu Stufe 5 oder höher	Überwinden von Fehlvorstellungen

Die Stufen werden folgendermaßen beschrieben:

- *Stufe 1 (Score 450): Lösen von Routineaufgaben mit Mittelstufenwissen*

  Auf dieser Stufe wird Wissen der Sekundarstufe I verlangt wie z.B. das Lesen eines Schaltdiagramms.

- *Stufe 2 (Score 550): Anwendung von Faktenwissen zur Erklärung einfacher Phänomene der Oberstufenphysik*

  Oberflächliche Kenntnisse von Konzepten und Gesetzen der Oberstufenphysik sind erforderlich, sie müssen aber nicht vertieft sein diese Stufe zu beschreiten.

- *Stufe 3 (Score 650): Anwendung physikalischer Gesetze (Größengleichungen) zur Erklärung experimenteller Effekte auf Oberstufenniveau*

Zur Lösung der Aufgaben dieser Stufe sind Kenntnisse über z.B. den Impulserhaltungssatz und die Fähigkeit, Gleichungen aufzustellen und aufzulösen erforderlich.

- *Stufe 4 (Score 750): selbstständiges fachliches Argumentieren und Problemlösen*

Stufe 4 verlangt eigenständige Lösungsansätze sowie die Fähigkeit, die Verbindung von Theorie zu Experiment herzustellen.

- *Stufe 5 (Score 850): Überwinden von Fehlvorstellungen*

Auf Stufe 5 gilt es, von zwar allgemeingültigen aber dennoch fehlerhaften Alltagsvorstellungen abzulassen. Eine Fehlvorstellung ist beispielsweise die Annahme, eine Welle bewege Material in Richtung ihrer Ausbreitung.

Schlussfolgerung

Aus der Betrachtung der TIMSS III wird deutlich, dass die Unterscheidung verschiedener Fähigkeitsniveaus die vorherige Analyse der Aufgabenanforderungen erforderlich macht. Nach theoretischen Überlegungen und empirischer Erprobung lässt sich dann sagen, welche Anforderungen ein Bearbeiter auf einer hohen, mittleren oder niedrigen Kompetenzstufe bewältigen kann. Durch verschiedene Studien wie z.B. TIMS III oder PISA liegen empirisch gestützte Kompetenzmodelle vor, die sich jedoch immer auf einzelne Bereicht und Altersgruppen beziehen. Klieme (2004) merkt an, die Beschreibung der Niveaus falle oft noch zu abstrakt aus. Vor allem fehle es auch an Kompetenzmodellen, welche die Entwicklung über die Jahrgangsstufen hinweg beschreiben konnten. Soziale oder interkulturelle Kompetenzen werden in den vorliegenden Modellen nicht behandelt. Möglicherweise ließen sich Kompetenzbereiche, die affektive Aspekte und Einstellungen einschließen, gar nicht klar voneinander abgrenzen geschweige denn auf einer Skala von „hoch“ bis „niedrig“ abbilden, so Klieme, sondern eher in Form von Mustern oder Typen darstellen. Die Anfänge der Kompetenzmodell- Entwicklung sind gemacht, die Weiterentwicklung von Messverfahren und deren Abnabelung von Fach- und Altersgruppengebundenheit steht noch aus.

Für den Bereich der Studierendenauswahl durch die Hochschulen ist zu überlegen, ob sich ein Kompetenzstufenmodell mit entsprechendem Test entwickeln ließe. Hierzu müsste vorab überlegt werden, welche Kompetenzen für zukünftige Studenten eines Faches notwendig sind. Darauf hin wäre ein Test zu entwickeln, der das Vorhandensein dieser Kompetenzen zuverlässig überprüfen kann.

Ein alternativer Weg wäre, die Fähigkeiten zu erheben, über die bereits erfolgreich studierende verfügen, um diese nach Übertragung in Aufgaben, deren Bearbeitung eben diese Fähigkeiten erfordern, in Testform für Studienbewerber umzuwandeln.

Die Entwicklung solcher Verfahren zur Studierendenauswahl wäre allerdings nur dann sinnvoll, wenn sich Fähigkeiten, die als Voraussetzung vorhanden sein müssen von solchen trennen ließen, die erst im Laufe des Studiums erworben werden.

Literaturverzeichnis

American Association for Advancement of Science (1994). Benchmarks for Science Literacy. Oxford University Press

Barnett, K. (1979). The Study of syntax variables. In G. A. Golding & C. E. McClintock (Hrsg.). Task variables in mathematical problem solving (S. 23-68). Philadelphia, PA: Franklin Institute Press

Bybee, R. W. (1997). Achieving Scientific Literacy: From Purpose to Practices. Greenwood Press

Klieme, E. (2004). Was sind Kompetenzen und wie lassen sie sich messen? In Pädagogik 06/2004

Klieme, E., Baumert, J., Köller, O. & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos & R. Lehmann (Hrsg.). TIMSS/ III. Dritte internationale Mathematik- und Naturwissenschaftsstudie. Band 1: Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit (S. 85-133). Opladen: Leske & Buderich.

Klieme, E. (2000). Fachleistungen im voruniversitären Mathematik- und Physikunterricht: Theoretische Konzepte, Kompetenzstufen und Unterrichtsschwerpunkte. In J. Baumert, W. Bos, & R. Lehmann (Hrsg.). TIMSS/III. Dritte internationale Mathematik- und Naturwissenschaftsstudie. Band 2: Mathematische und naturwissenschaftliche Grundbilung am Ende der gynmasialen Oberstufe (S. 57-128). Opladen: Leske & Buderich.

Lesgold, A., Lafoie, S., Logan, D. & Eggan, G. (1990). Applying cognitive task analysis and research methods to assessment. In N. Frederiksen, R. Glaser, A. Lesgold & M. G. Shafto (Hrsg.). Diagnostic monitoring of skill and knowledge acquisition (S. 325-391). Hillsdale, NJ: Erlbaum.

Moosbrugger, H. & Müller, H. (1997). Testmodelle der Item-Response-Theorie (IRT). Frankfurt am Main: Institut für Psychologie.

National Council of Teachers of Mathematics (NCTM). (1989). Curriculum and evaluation standards for school mathematics. Reston, VA: NCTM.

Reimann, P. & Chi, M. T. H. (1989). Human expertise. In K. J. Gilhooly et al. (Hrsg.). Human and machine problem solving (S. 161-191). New York: Plenum.

Shamos, M. H. (1995). The Myth of Scientific Literacy. Rutgers University Press.

Stein, M. K., Grover, B. W. & Henningsen, M. (1996). Building student capacity for mathematical thinking and reasoning: An analysis of mathematical tasks used in reform classrooms. American Educational Research Journal, 32 (2), (S. 455-488).

Stern, E. (1998). Die Entwicklung des mathematischen Verständnisses im Kindesalter. Lengerich: Pabst.

Steyer, R. & Eid, M. (2001). Messen und Testen. Berlin Springer

Williams, G. & Clark, D. (1997). Mathematical task complexity and task selection. In D. M. Clarke et al. (Hrsg.). Mathematics. Imagine the possibilities (S. 406-415). Brunswick, Victoria: Mathematics Association of Victoria.

Zur Übersicht und für Beispielaufgaben der TIMSS III:

http://www.timss.mpg.de

Berufserfolgsmessung

Micha Dombrowski

Einleitung

Berufserfolgsmessungen sollen wieder geben, wie erfolgreich eine Person in ihrem Beruf ist. Allerdings muss, wie für jede andere Messung zuvor genau definiert werden, was man messen möchte, das heißt in diesem Fall, was nun Erfolg ist und wie Erfolg sich zusammensetzt. Im Allgemeinen kann man sagen, der Erfolg entspricht der Leistung eines Mitarbeiters. Diese berufliche Leistung lässt sich als Beitrag zu den Zielen einer Organisation, also dem Unternehmen, definieren (Schuler, 2001, S.398). Somit kann der Mitarbeiter an dieser erbrachten Leistung gemessen werden, was für unsere leistungsorientierte Gesellschaft spricht. Im Wirtschaftsleben dominiert dieses Leistungsprinzip und wird auch als meritokratisches Prinzip bezeichnet.

Nun erfolgt das Messen der Leistung meist in der Form von subjektiven Beurteilungen. Dadurch, dass sie keinen passiven Messvorgang darstellt, wirkt sie ihrerseits als Intervention mit den verschiedensten Folgen. Leistungsbeurteilungen in leistungsorientierten Unternehmen sind ein gängiges Mittel, setzen aber gleichzeitig eine explizite Form oder einen formalisierten Ablauf nicht unbedingt voraus. Jedoch verfügt die Mehrzahl der deutschen Großunternehmen über formalisierte Beurteilungssysteme.

Diese sind meist standardisierte Formulare, mit denen der Vorgesetzte seine Mitarbeiter in regelmäßigen Abständen einschätzt. Dies wird von einem Mitarbeitergespräch gefolgt, welches auf beiden Seiten eher als unangenehm aufgefasst wird. Zum einen sieht sich der Mitarbeiter, bei ausgeübter Kritik des Vorgesetzten, schlechter beurteilt als er subjektiv seine eigene Leistung wahrnimmt, zum anderen kann der Vorgesetzte Bedenken über sein Ansehen bei den kritisierten Mitarbeitern haben. Schon allein an diesen beiden aufgeführten Problemen ist zu erkennen, dass Leistungsbeurteilungen nicht einfach durchzuführen sind. Allerdings ist der Nutzen eines solchen Instruments höher als die Risiken.

Neben der Leistung können auch andere Kriterien beruflichen Erfolgs, wie Berufs- und Arbeitszufriedenheit, Gesundheit und Wohlbefinden oder die eingenommene Position innerhalb der Organisation als Diagnoseziele formuliert werden. Allerdings wird auf diese Kriterien nicht weiter eingegangen.

Kriteriumsrelevanz, -defizienz und -kontamination

Was nun sind die Kriterien anhand derer der berufliche Erfolg festgemacht wird? Wenn man nun weiter von der beruflichen Leistung spricht und dies als Beitrag zu den Zielen einer Organisation definiert, so hat man nur eine sehr abstrakte Beschreibung des Ganzen.

Leistung ist als ein hypothetisches Konstrukt zu verstehen, dass als solches nicht direkt beobachtet werden kann (Schuler, 2001, S.399). Somit sind die Kriterien hierfür nur unvollkommene Annäherungen an dieses Konstrukt, da sie dien tatsächliche berufliche Leistung nur unvollständig abbilden und zusätzlich Irrelevantes mit erfassen.

Wenn wir von der Unvollständigkeit eines Kriteriums sprechen, so meinen wir damit die Defizienz. Wenn nun zusätzlich auch noch Irrelevantes mit erfasst wird, sprechen wir von Kriteriumskontamination. Beides findet sich in jedem Kriterium.

Ein Beispiel für Kriteriumsdefizienz kann man bei einem Außendienstmitarbeiter finden. Verkaufen zählt hierbei sicher zu den wichtigsten Aufgaben des Mitarbeiters und diese lassen sich einfach über die Umsatzzahlen erfassen. Also kann man die Umsatzzahlen als relevantes Kriterium für die Leitung des Mitarbeiters ansehen. Allerdings soll ein Außendienst-mitarbeiter nicht nur verkaufen, sondern auch Kunden an sich binden und die Konkurrenz beobachten. Über die Leistung auf diesen Gebieten geben die Umsatzzahlen keine Auskunft. Damit ist das Kriterium Unsatz in diesem Bereich defizient. Weiter kann man sagen, dass Umsatz nicht ausschließlich von der Leistung des Mitarbeiters beeinflusst wird. Verkaufaktivitäten der Konkurrenz und Konjunkturverlauf bestimmen den Verkaufserfolg mit und liegen außerhalb der Macht des Mitarbeiters. Damit ist das Kriterium Umsatz ebenso kontaminiert. Wie unten in der Abbildung 1 zu sehen ist, ist der Anteil der Relevanz relativ gering im Verhältnis zu Kontamination und Defizienz.

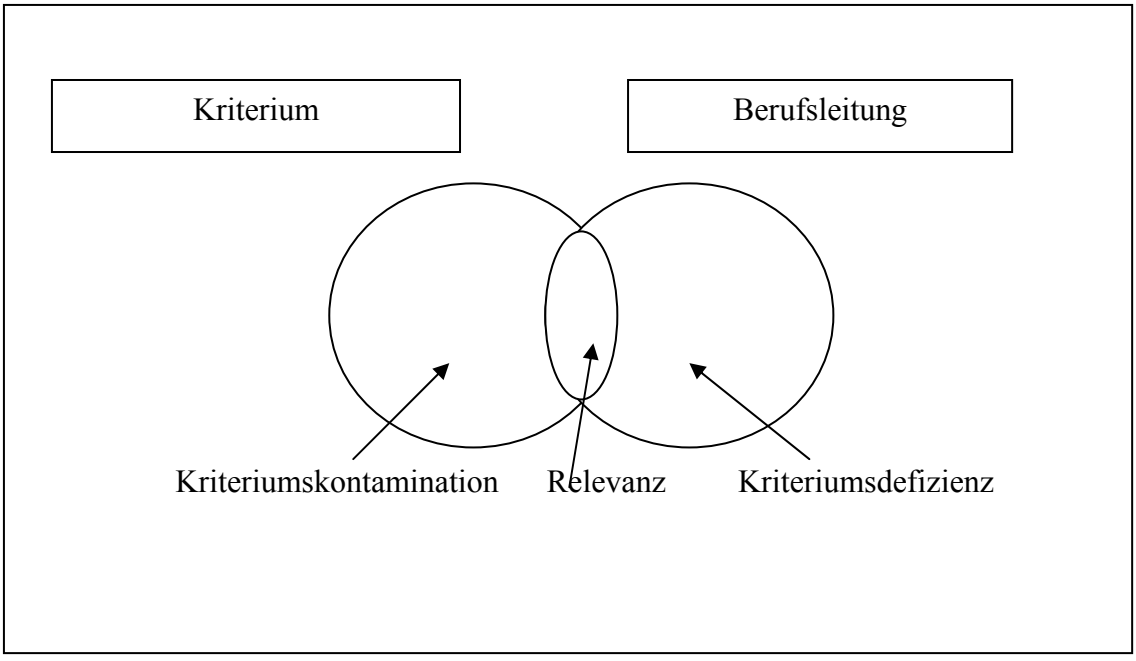


Abbildung 1: Kriteriumsrelevanz, -defizienz und –kontamination (Schuler, 2001, S.399)

Kenngrößen für Leistungserfassung

Nun lässt sich die Leistung am besten erfassen, indem man sich die Bereiche Potenzial, also Eigenschaften, Kenntnisse, Fähigkeiten usw., dann das Verhalten und die Ergebnisse einer zu beurteilenden Person genauer anschaut. Diese Dreiteilung wird häufig als Kausalkette aufgefasst, d. h. das eine bedingt das andere. Allerdings leidet dieses Theorem in der Praxis, da die Menschen nicht immer unter diesen drei Bereichen differenzieren können. Auch ist Objektivität sehr schwer einzuhalten, da sich hier das Problem der Defizienz und Kontamination bemerkbar macht.

Deshalb werden die Ergebnisse häufig nicht objektiv gemessen, sondern subjektiv beurteilt, womit sie anfällig für alle potentiellen Fehlerquellen des Urteilsprozesses werden. Nun gibt es verschiedene Kriterien, an denen man den Erfolg einer Person erkennen könnte. Durchgehend wichtig für jede Beurteilung, ob durch den Vorgesetzten oder durch andere Personen durchgeführt, sind Zielvereinbarungen. Diese Zielvereinbarungen werden prinzipiell mit dem Mitarbeiter und seinem Vorgesetzten vereinbart und nach Ablauf eines bestimmten Zeitraums wieder verglichen. Diese sollen dem Potential-, Verhalten- und Ergebnismöglichkeiten einer Person und den Vorstellungen eines Unternehmens entsprechen. Diese erfassen sozusagen den Istzustand und geben dann einen Sollzustand vor. Inhalte solcher Zielvereinbarungen können je nach Berufsparte unterschiedlich aussehen, da auch jedes Unternehmen unterschiedliche Ziele verfolgt.

Umsatz, Kundenkontakt und Teilnahme an Schulungen sind auf den ersten Blick, recht gute objektive Werte. Den erbrachten Umsatz einer Person kann man leicht erfassen auch nach Ablauf eines Jahres kann man sehr gut anhand der vorliegenden Umsatzzahlen diesen feststellen. Aber ein Problem bei diesem Maß ergibt sich darin, dass hier verschiedene Faktoren mitspielen und diese eben nicht so einfach ersichtlich sind wie Zahlen. Dazu gehören Faktoren außerhalb des Bereichs einer Person, wie Konjunkturverlauf, Qualität des zu vermarktenden Produkts oder Absatzpreis. Diese Faktoren haben ebenso einen starken Einfluss auf den Umsatz, aber kann von der Person nicht beeinflusst werden.

Auch ein auf den ersten Blick objektives Maß für den Erfolg einer Person im Beruf, ist deren Einkommen. Einfache Annahme wäre, je mehr eine Person verdient, desto mehr Erfolg hat sie im Beruf. Man geht davon aus, dass Unternehmen nur die Mitarbeiter honorieren, die besonders viel Leistung im Sinne des Unternehmens erbringen. Aber auch diese vermeintlich objektive Zahl ist nicht besonders gut geeignet, um den Erfolg auszumachen. Viele Berufssparten werden von Grund auf besser bezahlt, somit ergibt sich, dass eine Person die zwar über die Qualifikation verfügt, vielleicht gar nicht die gewünschte Leistung des Unternehmens erbringen kann, anfangs als gut bezahlter Mitarbeiter gilt.

Ebenso wichtig ist die Position, die ein Mitarbeiter innerhalb eines Unternehmens inne hat. Anhand derer könnte man ebenso den Erfolg festmachen. Zunächst müsste auch hier der Ausgangszustand festgehalten werden und ein Zeitraum, in der sich die Person weiterentwickeln kann, vereinbart werden. So könnte man den Anfangszustand, also die zu Beginn der Untersuchung besetzte Position, mit dem Endzustand, also eventuell erreichte Beförderung, nach abgelaufener Zeit vergleichen. Aber auch hier können sich Fehler einschleichen. Prinzipiell handelt es sich auf den ersten Blick um ein objektives Maß, aber bei genauerer Betrachtung kann man erkennen, dass die Zeit ein wesentliches Maß für das Kriterium Position ist. Eine Karriere innerhalb von fünf Jahren ist anders zu bewerten als innerhalb von 25 Jahren. Somit lässt sich erkennen, dass viele vermeintlich objektive Faktoren mit Fehlerquellen behaftet sind.

Wie man nun aus dieser Vielzahl von Möglichkeiten erkennen kann, ist Leistung kein einfaches Konstrukt. Es scheint sehr komplex und vielfältig zu sein und nicht auf alles und jeden übertragbar. Dennoch wurden erst in jüngster Zeit Versuche unternommen, generelle empirische und theoretische Aussagen über das Konstrukt beruflicher Leistung zu bekommen. Die Ergebnisse sprechen dafür, dass Leistung zwar mehrdimensional, aber nicht unüberschaubar ist, und somit das Ausmaß einer Generalisierbarkeit über verschiedene Arbeitsplätze enorm ist (Schuler, 2001, S. 402).

Quellen für Beurteilungen

Beurteilungen können auf verschiedene Arten erfolgen. Allerdings gilt für jede Art von Beurteilung, dass man prinzipiell sagen kann, menschliche Beurteiler sind grundsätzlich in der Lage, Defizienz und Kontamination bei objektiv erfassten Indikatoren mit zu berücksichtigen. Jedoch fließen situative Bedingungen stärker in subjektive Beurteilungen ein als in objektive Indikatoren. Zudem kommt hinzu, dass viele Leistungen sich nur durch Einschätzungen erheben lassen. Somit kann man sagen, dass eine der wichtigsten Quellen subjektiver Beurteilungen nach wie vor der direkte Vorgesetzte ist. Allerdings mangelt es den meisten Vorgesetzten an Gelegenheit die zu beurteilende Person in ihrem täglichen Arbeitsumfeld zu beobachten.

Eine weitere potentielle Urteilsquelle, die häufig Gelegenheit haben zur Beobachtung des Verhaltens einer zu beurteilenden Person, sind die unmittelbaren Arbeitskollegen bzw. Gleichgestellte. Dies hat den Vorteil, das Hemmschwellen und Leistungsablauf, wegen Beobachtung geringfügiger ausfallen als bei der Beurteilung durch den Vorgesetzten. Jedoch ist diese Beurteilung ebenfalls nicht fehlerfrei, denn hier wird der Einfluss von Sympathie oder auch Antipathie vermutet. Sicherlich fällt es einer Person leichter einen Kollegen, den man sympathisch findet besser zu beurteilen als einen den man weniger leiden kann. Allerdings greift dieser Einwand auch bei der Beurteilung durch andere Personen und ist nicht nur ausschließlich unter Gleichgestellten zu finden.

Eine dritte Quelle, die in letzter Zeit mehr Aufmerksamkeit geschenkt wird, sind Beurteilungen von unterstellten Mitarbeitern. Der Vorteil liegt hier in den unterschiedlichen Perspektiven, wobei der Fokus weniger auf Ergebnisse und Sachaufgabenerfüllung als auf interpersonale Aspekte der Mitarbeiter Führung gerichtet ist.

Auch hier muss erwähnt werden, dass die Qualität der Beurteilung, sowie die Bereitschaft überhaupt eine abzugeben, mit der Anonymität steht und fällt. Die Quelle mit konkurrenzlos direktem und umfassendem Zugang zum tatsächlichen Verhalten ist der Beobachtende selbst. Allerdings ist diese Beurteilung besonders anfällig für zufällige oder absichtliche Verzerrungen, was sich besonders an übermäßig vorteilhaften Urteilen bemerkbar macht. Diesen kann allerdings entgegengewirkt werden, indem man vorher mitteilt, dass die Urteile überprüft werden können.

Messungen und deren Skalenformate

Messungen erfordern Messinstrumente, so auch Messungen der beruflichen Leistung. Die meist verwendeten Skalenformate sind die der Einstufungsverfahren. Hier werden Ausprägungen von Merkmalen auf mehrstufigen Skalen eingeschätzt. Der Vergleich zwischen verschiedenen Personen findet erst auf einer weiteren Stufe statt. Für die Abstände zwischen den Skalenpunkten wird meist ein metrisches Skalenniveau vorausgesetzt, so dass statistische Verfahren angewendet werden können. Der einfachste Fall ist die Graphische Einstufungsskala, hier findet die Umsetzung als relevant erachteter Merkmale in Skalen ohne ein formales Skalierungsverfahren statt. Verhaltensverankerte Einstufungsskala, auch BARS (Behaviorally Anchored Rating Scales) genannt, werden nach einer komplexen mehrstufigen, inzwischen mehrfach revidierten Schema konstruiert, dessen Grundprinzip, an der Einstellungsmessung bekannten Thurstone-Skalierung (Schuler, 2001, S. 410) angelehnt ist. Als erstes werden von einer Gruppe von Beurteilern relevante Leistungsdimensionen definiert. Eine zweite Gruppe formuliert dann konkrete Verhaltensbeispiele für jede Dimension. Diese Beispiele werden von einer dritten Gruppe in Dimensionen zurückübersetzt. Die nun vorhandenen Beispiele werden von einer weiteren Gruppe aus dieser Stichprobe den einzelnen Skalenstufen zugeordnet. In der unten aufgeführten Abbildung 2 sind einige Beispiele für verschiedene Skalenformate aufgeführt.

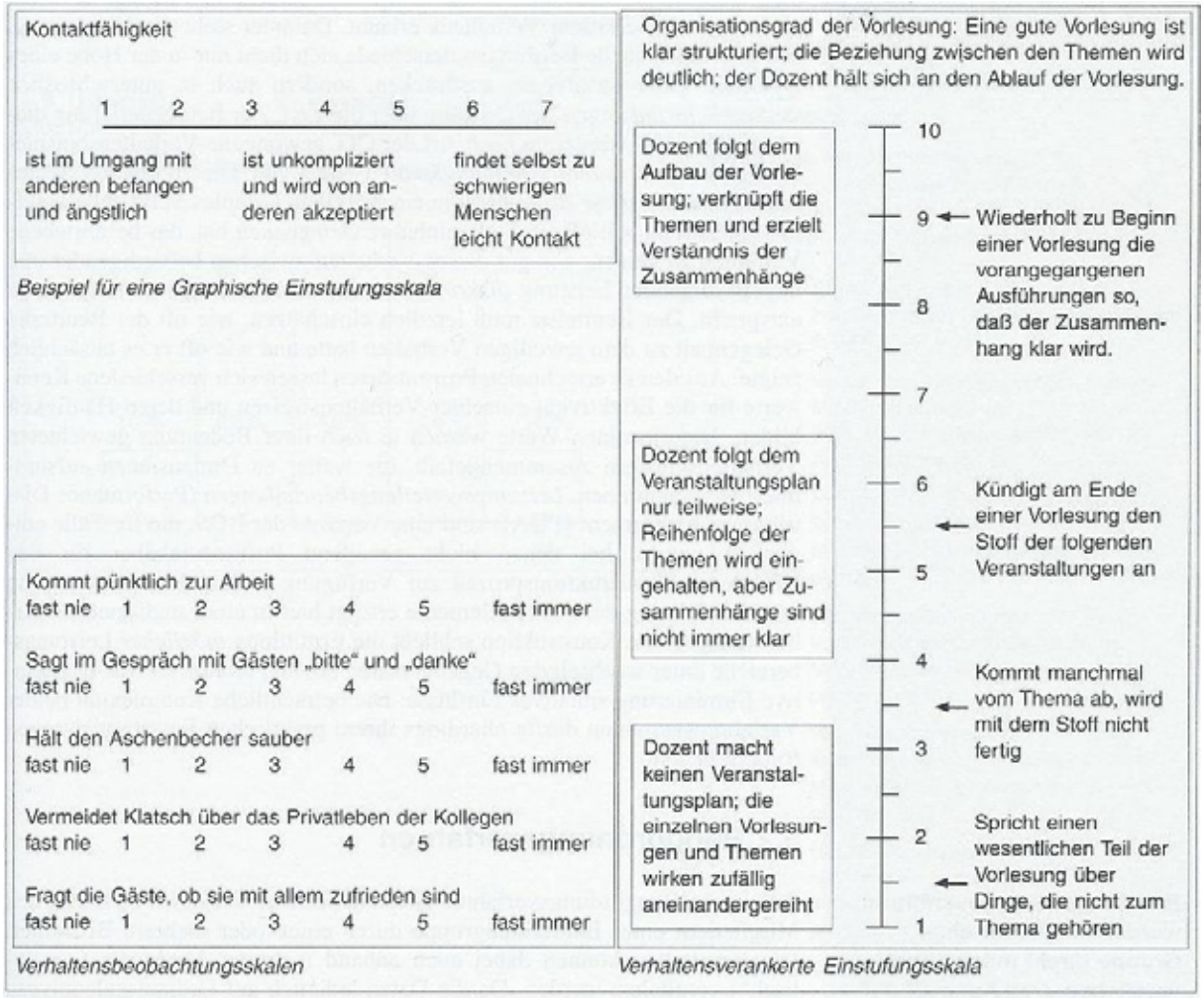


Abbildung 2: Beispiele für Einstufungsverfahren (Schuler, 2001, S.411)

Die BARS ist wohl auch das am besten erforschte Skalenformat. MSS (Mixed Standard Scale), die auf die Finnen Blanz zurückgeht, ist ähnlich der BARS, allerdings endet diese in einem anderen Format, welches eher der Guttman-Skala entspricht. Die Verhaltensaussagen werden kumulativ geordnet. Hier werden die Beurteilten danach eingestuft, ob sie besser, schlechter oder gleich gut sind, wie das formulierte Item vorgibt. Zu den neueren und weniger gut erforschten Einstufungsverfahren gehört das Modell der Verteilungsmessung. Diese kann eine Aussage über die Konsistenz des erfassten Verhaltens machen. Hier geht man davon aus, dass Leistungsunterschiede nicht nur auf unterschiedliches Leistungsniveau zurückzuführen ist, sondern auch auf unterschiedlich starken Schwankungen der Leistung über die Zeit. Es wird somit erfasst, wie oft der zu Beurteilende die Gelegenheit hat, das beschriebene Verhalten zu zeigen und ob er es auch wirklich gezeigt hat.

Rangordnungsverfahren

Bei Rangordnungsverfahren werden die Mitglieder einer bestimmten Gruppe direkt nebeneinander verglichen. Da die Daten hier ausschließlich ordinalskaliert sind, lässt sich keine Aussage machen über das Ausmaß der Leistung, sondern es wird ausschließlich

differenziert zwischen den einzelnen Gruppenmitgliedern. Im einfachsten Fall werden direkte Rangreihen auf der Grundlage globaler oder differenzierter Kriterien gebildet. Typischerweise basiert die Rangreihenbildung auf der so genannten Quotenvorgabe. Hierbei werden die Beurteiler aufgefordert extreme Rangplätze seltener zu vergeben, so dass die Verteilung einer Normalverteilung annähert. In der nachfolgenden Abbildung 3 sind einige Beispiele aufgeführt für Rangordnungsverfahren.

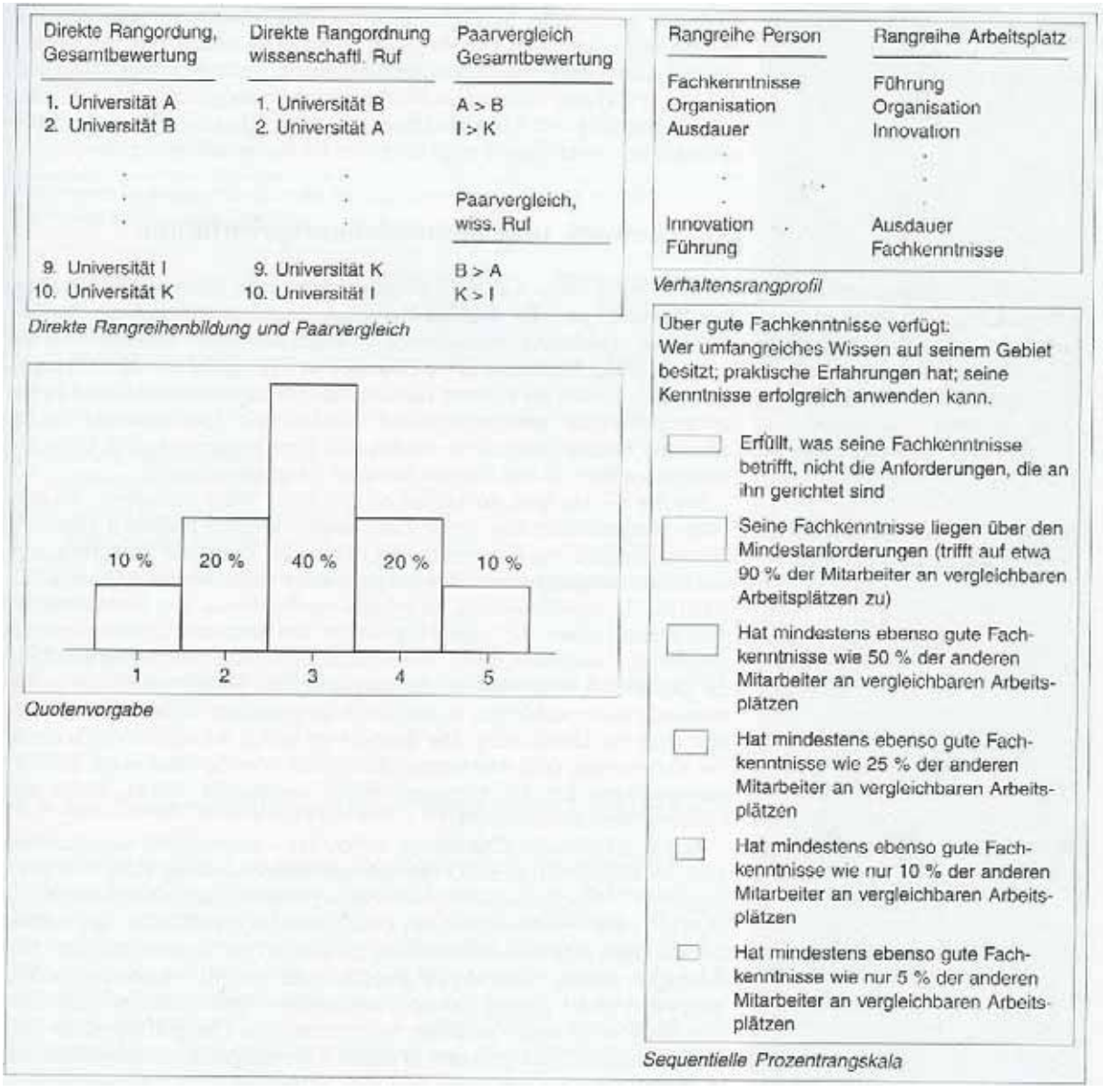


Abbildung 3: Beispiele für Rangordnungsverfahren (Schuler, 2001, S.413)

Auswahl- und Kennzeichnungsverfahren

Bei Auswahl- und Kennzeichnungsverfahren wird die Zustimmung oder Ablehnung vorgegebener Aussagen oder aus einer Liste nach Multiple-Choice-Prinzip von den Beurteilern verlangt. Dadurch dass den Beurteilern unbekannt ist, wie günstig ein Urteil ist,

kann verschiedenen Urteilstendenzen entgegengewirkt werden. Diese Verfahren zählen auch zu den ältesten formalen Beurteilungsskalen. Eine Form hierfür ist die Auswahlliste mit freier Wahl. Diese wird nach dem Thurstone-Verfahren aufgebaut bis schließlich eine Aussageliste verbleibt, die nach der Möglichkeit den gesamten Leistungsbereich gleichmäßig abdeckt. Hier dient der Median der Effektivitätsgewichte für die Aussagen, denen zugestimmt wurde, als Leistungswert des Beurteilten. Anders funktioniert das Wahlzwangsverfahren. Dieses verlangt von Beurteilern aus der Liste von jeweils zwei oder mehreren Aussagen diejenige auszuwählen, die ihnen am treffendsten erscheinen.

Evaluation

Wenn verschiedene Messinstrumente desselben Selektionsprozesses systematisch auf ihren Betrag zur Vorhersage des Berufserfolges überprüft und hieraus entsprechende Konsequenzen gezogen werden, kann man von einer Evaluation sprechen.

Um die verschiedenen Instrumente nun miteinander vergleichen zu können, müssen sie in einigen Punkten übereinstimmen. Das heißt es müssen dieselben Bedingungen herrschen, damit miteinander verglichen werden kann.

Reliabilität und Validität

Wie man sehen konnte, werden Leistungen mit unterschiedlichen Kriterien, Quellen und Verfahren beurteilt. Die Evaluation diese Vorgehensweisen erfordert ihrerseits eine Beurteilung anhand von Kriterien für Kriterien, die sich noch mal einander messen lassen.

Ein nahe liegendes Vorgehen zur Evaluation von Berufserfolg ist der Rückgriff auf die Klassische Testtheorie. Allerdings erfordert die Übertragung einige Anpassungen. Interne Konsistenz ist bei Berufsleistung nicht sinnvoll anwendbar, ebenso wenig wie die Parallelitätsreliabilität. Allerdings kann die Retestreliabilität recht gut ermittelt werden.

Noch schwieriger als die Überprüfung der Reliabilität gestaltet sich Überprüfung der Validität bei Leistungsbeurteilungen, dies liegt vor allem an der nicht hinreichenden Definition des Konstruktes für Berufsleistung bzw. Berufserfolg.

Genauigkeit

Genauigkeit ist im Gegensatz zu Validität und Reliabilität ein Konzept, dass sich sehr gut für Leistungsbeurteilungen anwenden lässt. Genauigkeit, auch Akkuratheit (accuracy) eines Urteils ist definiert, als Abweichung von einem wahren Leistungswert. Problem der Akkuratheit ist, dass sie sich unter Laborbedingungen als ein höchst genaues Maß bewährt hat, aber auch genau dies ist der Einwand für dieses Kriterium, da es bisher ausschließlich in Laborexperimenten bewährt hat.

Praktikabilität

Eines der größten Probleme ergibt sich in der Anwendung der verschiedenen Verfahren zur Messung von Berufserfolg. Dieser Punkt wird bei vielen wissenschaftlichen Arbeiten kaum berücksichtigt. Viele Beurteilungssysteme benötigen aufwendiges Training der Beurteiler, was in vielen Situationen nicht realisierbar ist. Somit ist die Umsetzung eines Beurteilungsverfahrens in einem Unternehmen von der Einfachheit in der Anwendung stark abhängig.

Akzeptabilität

Stärker noch als die Praktikabilität ist die Akzeptanz eines Verfahrens. Beides sind nicht hinreichende Bedingungen, aber ausgesprochen notwenige. Die Akzeptanz hängt mitunter vor der Gerechtigkeit eines solchen Beurteilungssystems ab. Wenn diese als unfair empfunden werden, kann das bei den Beteiligten zu ernsten negativen Reaktionen führen.

Fazit

Nach Lany und Farr kam man zu dem desillusionierenden Schluss, dass kein entscheidender Durchbruch zu erkennen war, Leistungsbeurteilungen durch verfeinerte Skalenformate zu verbessern (Schuler, 2001, S.420). Die unten aufgeführte Tabelle 1 stelle einen Versuch dar, vergleichende Untersuchungen zur relativen Eignung verschiedener Beurteilungsverfahren in maximaler Weise zu verdichten.

Natürlich ist diese Tabelle nur eine starke Vereinfachung, doch lässt sich daraus erkennen, dass Verfahren, deren innere Logik den Beurteilern unbekannt ist, für Feedback weniger geeignet sind. Dies gilt auch für Akzeptabilität. Sie ist nur eine einfache Möglichkeit, die verschiedenen Verfahren miteinander zu vergleichen.

Tabelle 1: Evaluation verschiedener Beurteilungsverfahren (Schuler, 2001, S.421)

	Reliabilität	Akkuratheit	Praktikalibilität	Akzeptabilität	Verh.strg.	adm. Entsch
Graph. Skala	0	-	+	0	0	0
BARS	0	0	0	+	+	0
BOS	0	0	+	0	+	0
MSS	0		+	0	-	0
PDA/BDS	-	0			-	0
Direkte Rangreihenbildung	+		0	-	--	(+)*
Paarvergleich	++		+	-	--	(+)*
Verhaltens-Rangprofil	++		0		++	(--)**
Checklist	0		+	-	-	0
Wahlzwangsv.	+		-	--	--	+

*= nur für die Laufbahnentsch. für Entgeltfindung ungeeignet

**= o. zusätzliche Verwendung einer Einstufungsskala

fünfstufige Qualitätsskala mit Werten von ++, +, 0, -, --

Praktische Aspekte

Festzuhalten ist, dass die Einführung formaler Beurteilungen für Erfolg ein komplexes, innovatives Unterfangen ist, in dessen Verlauf personelle und sachliche Ressourcen wechselnder Quantität und Qualität gebunden werden. Nun kann man nicht nur Berufserfolg auf diese Weise erfassen, sondern auch Studienerfolg. Natürlich müssten hier die Kriterien angepasst werden und entsprechend in die Beurteilungsverfahren mit einfließen. Beurteilungen durch Gleichgestellte würden hier nicht eine so gewichtete Rolle spielen, da nach Abschluss eines Studiums die Wege auseinander laufen. Aber Beurteilungen durch Vorgesetzte, also Professoren, ist auch nicht der ideale Weg, da wie bei einem Unternehmen, dieser nicht besonders gut und ausreichend einen Studenten beobachten kann. Eine gute Alternative wäre hier ein Mentor, der den Student durch sein

Studium hinweg begleitet und regelmäßig, vielleicht durch Eigenbeurteilungen und Beurteilungen der Professoren, sich ein Bild über den Erfolg des Studiums machen kann.

Messung von Studienerfolg

Nun kann Studienerfolg nicht ausschließlich durch solche Beurteilungen ausgemacht werden. Ein recht großer Unterschied besteht in der Auffassung von Zeit. Während eine lange Karriere im Berufsleben erstrebenswert ist, so sollte der Aufenthalt an einer Universität bis zum Abschluss recht kurz sein, d.h. ein schneller Abschluss kann ein gutes Zeichen für Erfolg sein. Längs- und Querschnittstudien sind hier ebenso von Nöten wie bei Berufserfolg. Der Begriff Leistung, der zuvor mit Erfolg gleichgesetzt worden ist, sollte auch neu definiert werden bei Studienerfolg, denn diese ist während eines Studiums anders als im Beruf. Während im Beruf Bindung von Kunden oder anderen Geschäftkontakten leistungsfördernd ist, so entspricht das nicht während eines Studiums, da hier der Nutzen nicht gleich groß ist. So ist es zwar angenehm, sich mit Kommilitonen zusammen zu tun und vielleicht auch hilfreich, doch letztendlich, nimmt dies nicht einen so großen Einfluss zum Beispiel auf die Abschlussnoten. Allerdings ist Studienerfolg ein gutes Kriterium für die Vorhersage von Berufserfolg, auch wenn nicht unfehlbar und geschlechtsübergreifend.

Geschlechtsspezifische Unterschiede im Berufseinstieg

In den ersten und zweiten Befunden der Erlanger Längsschnittstudie, die sich mit der Vorhersage von Berufserfolg von Hochschulabsolventinnen und -absolventen beschäftigt, hat sich ein Unterschied zwischen den Geschlechtern gezeigt, was ihren weiteren Werdegang betrifft (Abele-Brehm, A. & Stiel, M., 2004). Nun konnte aber man an allen vorgestellten Verfahren sehen, dass eigentlich kein Unterschied zwischen Mann und Frau gemacht wird. Dennoch gibt es einen Unterschied im Einstieg ins Berufsleben. Dieser ist nicht im Ergebnis zu sehen, denn da schneiden Männer wie Frauen gleich gut ab. Dennoch schaffen Frauen nicht einen gleich guten Einstieg in das Berufsleben wie Männer, auch wenn der Studienerfolg gleich ist. So lässt sich festhalten, dass Frauen auch weniger Erfolg im Beruf haben als Männer. Aber wie kommt es, dass Frauen schon tiefer in einen Beruf einsteigen als Männer?

Dies kann auf viele verschiedene Faktoren zurückgeführt werden. Einer der Faktoren ist das Alter. Frauen die erfolgreich ihr Studium abgeschlossen haben, werden auf den Berufsmarkt in einem Alter geworfen, in denen sie von vielen Arbeitgebern, als Risiko betrachtet werden. Dieses Risiko besteht darin, das der weibliche Mitarbeiter schwanger werden könnte. Das Problem an sich ist für den Arbeitgeber nicht unbedingt die Schwangerschaft, doch der damit verbundene Ausfall. Ein Arbeitgeber möchte keine Stelle kurzfristig besetzen, er möchte für sich den maximalen Erfolg aus einem Mitarbeiter ziehen und bei Möglichkeit so lange wie möglich. So befürchtet nun das Unternehmen, wenn es Frauen mit Führungspositionen betreut, in dem vermeintlichen Risikoalter zwischen 25-35

Jahren, wobei auch hier eine Verschiebung des Alters nach hinten stattfindet, dass Verlust machen könnte. Zwar scheidet die Mitarbeiterin nicht aus, dennoch steht sie nicht im vollen Umfang mehr zu Verfügung. Somit würde es erklären, warum man Frauen eher etwas tiefer beruflich einstuft werden als Männer, bei gleicher Ausgangsposition.

Man könnte auch das klassische Missverhältnis zwischen den Geschlechtern in den Unternehmen anführen, warum Frauen mit gleichem Studienerfolg schlechter bewertet werden, aber wir hoffen doch sehr, dass die Zeit der Unterdrückung des weiblichen Geschlechts vorbei ist. Ein anderer interessanter Aspekt ist, das Frauen mehr im Team leisten als Männer, diese aber besser als Einzelkämpfer Leistung bringen. Man könnte nun den Unternehmen unterstellen, dass sie dieses für sich ausnutzen und Frauen eher in teamorientierten Positionen sehen und Männer in Führungspositionen. Das Problem hierbei ist, dass dies vielleicht nicht falsch ist, aber sich der gleiche Studienerfolg immer noch nicht erklären lässt. So müsste eine Frau auch im Studienerfolg schlechter abschneiden, was sie aber nicht tut. Diese Unterschiede zwischen Männer und Frauen im Berufserfolg lässt sich nicht mit Studienerfolg erklären, da wir zuvor erwähnt die Ausgangsbasis gleich ist

Fazit

Wie man aus den vorhergehenden Anschnitten ersehen konnte, kann Berufserfolg gemessen werden, jedoch nicht so einfach. Es gibt viele Kriterien die den Erfolg im Beruf ausmachen und es ist schwer diese einfach zu identifizieren und zu messen, da zumal viele Kriterien subjektiv erfasst werden, nämlich durch Beurteilungen. Diese Verfahren lassen sich in abgewandelter Form auch für die Messung von Studienerfolg anwenden. Wobei hier wichtig ist, dass die Leistung im Beruf nicht die gleiche ist, wie im Studium. Was besonders interessant ist, ist der Zusammenhang von Studienerfolg und Berufserfolg. Durch den Studienerfolg kann man recht gut den Berufserfolg vorhersagen. Diese signifikanten Zusammenhänge könnten in den verschiedensten Bereichen genutzt werden So könnte man zum Beispiel schon bei einer Einschätzung des Erfolgs einer Person in einem Studium auch die Wahrscheinlichkeit für den Erfolg im späteren Beruf machen. Dies hätte nicht nur Vorteile für die Wirtschaft, sondern auch für Hochschulen. Eine entsprechende Studierendenauswahl kann hier von Vorteil sein. Diesen Bewerbern könnte von Studiengängen abgeraten und stattdessen günstigere Fächer aufgezeigt werden, in denen der Bewerber wahrscheinlich einen besseren Abschluss machen würde und auch später bessere berufliche Aussichten hätten. Kürzere Studienzeiten und bessere Abschlussnoten könnte die Folge sein. Für die entsprechende Auswahl von Studierenden müsste allerdings die Methode der Berufserfolgsmessung, nicht erst nach dem Studium ansetzten, sondern schon weitaus früher. Durch die Messung des Schulerfolgs auf der Basis der Messung des Berufserfolges könnte so die Qualifikation einer Person für ein Studium erhoben werden. Anzumerken wäre hier wie schon bei Berufserfolg, Erfolg lässt sich nicht ausschließlich an Zahlen festhalten, also Noten. Es müssten weitere wichtige Kriterien berücksichtigt werden,

die für ein erfolgreiches Studium wichtig sind. Allerdings sind solche Erhebungen nicht einfach zu realisieren. Bisher wurde in Deutschland nur ein Bruchteil der Kriterien für Studienerfolg berücksichtigt.

In einigen Studiengängen, wie Medizin, in denen mehrer dieser Kriterien eine Rolle spielten, wurde sogar wieder verzichtet mit dem Argument aus Kostengründen und vermeintlicher Chancengleichheit. Um alle Kriterien ausreichend zu berücksichtigen müsste ebenso der jährliche Aufwand für die zu modifizierenden Verfahren sowie Testdurchführung und anschließende Auswertung in Rechnung gestellt werden (Rindermann, H. & Oubaid, V. ,1999). Dennoch wäre dies ein interessanter Aspekt, da auch die Wirtschaft von solchen Evaluationen ihren Nutzen ziehen kann. Sie müsste nicht erst warten bis ein geeigneter Kandidat auf den Markt kommt und ihn anhand seiner Abschlussnoten bewerten. Unternehmen könnten anhand der erhobenen Wahrscheinlichkeit des Studiumserfolgs schon frühzeitig geeignete Bewerber angehen.

Literaturverzeichnis

Abele-Brehm, A. & Stiel, M. (2004). Die Prognose des Berufserfolgs von Hochschulabsolventinnen und –absolventen. Befunde zur ersten und zweiten Erhebung der Erlanger Längsschnittstudie BELA-E. *Zeitschrift für Arbeits- und Organisationspsychologie*, 48, 4-16.

Höft, S. (2001). Erfolgsüberprüfung personalpsychologischer Arbeit. In: H. Schuler (Hrsg.). *Lehrbuch der Personalpsychologie* (S.618-648). Göttingen: Hogrefe

Marcus, B. & Schuler, H. (2001). Leistungsbeurteilung. In H. Schuler (Hrsg.), Lehrbuch der Personalpsychologie (S. 397-431). Göttingen: Hogrefe.

Schuler, H. & Funke, U. (1995). Diagnose beruflicher Eignung und Leistung. In Schuler (Hrsg.), *Lehrbuch Organisationspsychologie* (2., Korr. Aufl.) (S.235-283). Bern: Huber.

Kelloway, E.K. (1998). *Using LISREL for Structural Equation Modeling*. Thousands Oak: Sage.

Rindermann, H. & Oubaid, V. (1999). Auswahl von Studienanfängern durch Universitäten – Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. *Zeitschrift für Differenzielle und Diagnostische Psychologie*, 20, 172-191.

Anforderungsanalyse

Nina Roczen

Zur Notwendigkeit einer Anforderungsanalyse bei der Studierendenauswahl durch Hochschulen

Eine Anforderungsanalyse sucht nach Voraussetzungen auf Seiten der Person, die diese zur Bewältigung einer bestimmten Tätigkeit in einem bestimmten Kontext befähigen (Jeserich, 1981). So ist es auch bei der Entwicklung eines universitären Auswahlverfahrens von großer Wichtigkeit, die Voraussetzungen, die Studenten zum erfolgreichen Bewältigen eines bestimmten Studienganges mitbringen müssen, zu ermitteln, indem die Anforderungen eines spezifischen Studienganges, Studienerfolgskriterien und prognostisch relevante Eingangsmerkmale auf Seiten der Studienanfänger bestimmt werden (Amelang 1997, Viek et al. 1997, Rindermann & Oubaid, 1999).

Ein Studieneingangstest dient der Beurteilung der Passung zwischen Student und Studiengang. Eine vorangestellte Anforderungsanalyse stellt dabei die qualitätssichernde Basis dieser Beurteilung dar.

Bestimmung der Anforderungen

Um die angestrebte Passung zwischen Student und Studiengang zu erreichen, muss ein Vergleich zwischen Student und Studiengang auf verschiedenen Ebenen stattfinden. Neben den *tätigkeitsspezifischen Anforderungen*, d.h. Fähigkeiten, Fertigkeiten und Kenntnisse, die erforderlich sind, um Leistungen am Arbeitsplatz bzw. in einem bestimmten Studienfach zu erbringen, sollten auch *tätigkeitsübergreifende Anforderungen* bestimmt werden. Diese sind insofern bedeutsam, als sich nicht alle eignungsrelevanten Anforderungen einer Tätigkeit ermitteln lassen, da sie sich in zum Teil nicht vorhersehbarer Weise verändern. Tätigkeitsübergreifende Aspekte fordern deshalb auch Flexibilität, Lern- und Entwicklungsfähigkeit, um sich veränderlichen Anforderungen adäquat anpassen zu können. Schließlich spielt auch das *Befriedigungspotential* der zur Frage stehenden Tätigkeit mit Interessen, Werthaltungen und Bedürfnissen eine Rolle, um Zufriedenheit, Wohlbefinden sowie Bindung an einen Studiengang zu gewährleisten, um damit das Risiko eines Studienabbruchs zu verringern (vgl. Schuler, 2001).

Grundsätzliche Zugänge zur Anforderungsbestimmung

Schuler (2001) unterscheidet die drei folgenden grundsätzlichen Möglichkeiten zur Bestimmung von Anforderungen:

Die *erfahrungsgeleitet – intuitive Methode* beschäftigt sich mit den Eigentümlichkeiten der Berufe bzw. Studiengänge, d.h. mit den dabei auszuübenden Tätigkeiten sowie mit den dortigen Gegebenheiten und Bedingungen. Dies erfolgt ohne den Einsatz empirischer Methoden und verlangt daher vom Anwender ein hohes Maß an Erfahrung. Die erfahrungsgeleitet – intuitive Methode kommt vor allem im Rahmen der Berufsberatung zum Einsatz.

Mit der *arbeitsplatzanalytisch-empirischen Methode* werden die Tätigkeiten und Situationen eines konkreten Arbeitsplatzes bzw. eines Studienganges mit Hilfe von Fragebögen untersucht. Anschließend können die auf diese Weise gewonnenen Tätigkeitsmerkmale über einen Arbeitsplatz zusammengefasst und anschließend in zu einer erfolgreichen Bewältigung dieser Tätigkeit erforderliche Personmerkmale übersetzt werden. Für die meisten Fälle bezeichnet Schuler (2001) diese Vorgehensweise als die Methode der Wahl.

Die *personbezogen-empirische Methode* versucht, über statistische Zusammenhänge zwischen den Merkmalen der in einem Beruf tätigen Personen einerseits und Kriterien wie Leistungshöhe und Berufszufriedenheit andererseits die Anforderungen an die Person sowie die Befriedigungsangebote der Tätigkeit zu bestimmen. Die Anwendung dieser Methode ist dann nicht indiziert, wenn es sich um Personmerkmale handelt, die in starkem Maße durch Übung und Training beeinflusst werden können, wie Kenntnisse und Fähigkeiten. Weniger gravierend ist dieses Problem im Falle von Fähigkeiten oder Temperamentseigenschaften.

Schuler und Funke (1995) schlagen vor, bei der Bezeichnung der oben genannten Verfahren den Terminus *Arbeitsanalyse* zu verwenden, sofern ein Arbeitsplatz in Situationsbegriffen beschrieben wird, bei der Beschreibung in Personbegriffen hingegen von *Anforderungsanalyse* zu sprechen. So würde man bei der arbeitsanalytisch-empirischen Methode zunächst von ‚Arbeitsanalyse', nach der Übersetzung der Tätigkeits- in Personenmerkmale von ‚Anforderungsanalyse' sprechen. Bei den meisten Autoren im deutschsprachigen Raum werden die Begriffe entweder gemeinsam (‚Arbeits- und Anforderungsanalyse') oder synonym verwendet.

Methoden der empirischen Anforderungsanalyse – Klassifikationsversuche

Für die Beschreibung der einzelnen Methoden der Arbeits- und Anforderungsanalyse wurden verschiedene Möglichkeiten der Klassifikation vorgeschlagen:

Fleishmann und Quaintance (1984) wählten hierfür die Unterscheidung zwischen Arbeits- und Anforderungsanalyse wie oben erläutert und rechneten lediglich Aufgabenbeschreibungen der Arbeitsanalyse, die behavior description - Ansätze

(Verhaltensbeschreibung), behavior requirement - Ansätze (Verhaltenserfordernisse) und die ability requirement - Ansätze (Fähigkeitserfordernisse) der Anforderungsanalyse zu.

Eine weitere Klassifikation der Arbeits- und Anforderungsanalyse ist die nach den Quellen der arbeitsanalytischen Information (Schuler, 2001). Hierbei wird ein Verfahren danach kategorisiert, wer der Durchführende der Analyse (Arbeitsplatzinhaber bzw. Student, Vorgesetzter bzw. Dozent oder Arbeitsanalytiker) ist und welche Analysemethoden (Beobachtung, Interview, Fragebogenerhebung, Untersuchung des Arbeitsmaterials) eingesetzt werden.

Für die im Folgenden beschriebenen Methoden der Arbeits- und Anforderungsanalyse wird die Klassifikation ‚aufgabenbezogen', ‚verhaltensbezogen' und ‚eigenschaftsbezogen' verwendet.

Die Beschreibungsebenen Aufgaben-, Verhaltens- und Eigenschaftsbezug

Die Differenzierung dieser drei Beschreibungsebenen hat sich vor allem im personalpsychologischen Bereich als eine praktisch sehr hilfreiche Unterscheidung erwiesen, denn sie bietet sich zur Zuordnung entsprechender eignungsdiagnostischer Verfahren und Leistungskriterien an (Schuler 1989, Schuler & Funke, 1995). Auswahlverfahren und Personalentwicklungsinstrumente können, wie in Abb.1 veranschaulicht, auf diese Weise unmittelbar auf die Anforderungsanalyse bezogen und durch diese begründet werden. Es ist zu vermuten, dass die Einhaltung der gleichen Ebene für Anforderungsbestimmung, Eignungsdiagnostik und Personalentwicklung zu nachvollziehbaren und letztlich erfolgreicheren Maßnahmen führt (Schuler, 2001).

Arbeits- und Anforderungsanalyse	Eignungsdiagnostisches Auswahlverfahren	Maßnahmen der Personalentwicklung	Leistungskriterien
Aufgaben-, Ergebnis- und Qualifikationsanforderungen	Kenntnisprüfungen, Noten, Biographie, fachliche Qualifikation und Erfahrung	Wissensorientierte Verfahren, Bildung, fachliche Qualifizierung	Ergebniskriterien, Qualitätskriterien, Standards, Examina, Erledigung, Zielerfüllungsgrad
Verhaltensanforderungen, z.B. Fertigkeiten, Gewohnheiten, Handlungsregulation	Arbeitsproben, Simulationen, Fertigkeitsprüfungen	Verhaltensorientierte Verfahren, stellenbezogene Entwicklung, Coaching	Verhaltenskriterien
Eigenschaftsanforderungen, z.B. Fähigkeiten, Temperamentsmerkmale, Interessen	Tests, Potentialanalyse	Persönlichkeitsentwicklung, Sozialisation	Eigenschaftskriterien

Abb. 1. Beschreibungsebenen personalpsychologischer Instrumente (Schuler, 2001)

Aufgabenebene

Bei der aufgabenorientierten Arbeitsanalyse werden die objektiven Tätigkeiten einer Stelle oder eines Studiengangs beschrieben, dies erfolgt zumeist mittels sogenannter Aufgabeninventare, anhand deren Items die Bedeutung, die Schwierigkeit sowie die Häufigkeit der betreffenden Aufgabe beurteilt werden kann. Je nachdem, wie detailliert die Beschreibung einer Tätigkeit erfolgen soll, ist ein speziell für die Tätigkeit entwickeltes Inventar zu verwenden. Für eignungsdiagnostische Zwecke ist eine Bestimmung der eher allgemeinen, typischen Charakteristika einer Tätigkeit sinnvoll, für Trainingszwecke kann eine sehr detaillierte Beschreibung hilfreich sein. Ein Beispiel für eine aufgabenorientierte Arbeitsanalyse ist die Hierarchische Aufgabenanalyse (Annett & Duncan, 1967), die auf der Annahme basiert, dass Verhalten hierarchisch organisiert ist und Analysekategorien wie„Pläne“ und „Operationen“ verwendet.

Verhaltensebene

Analysen auf der Verhaltensebene bilden die gängigste und methodisch vielfältigste Form der Anforderungsanalyse.

Ein standardisiertes Verfahren zur arbeitsplatzübergreifenden Beschreibung des Verhaltens (obwohl einige Items auch aufgaben- oder eigenschaftsbezogen formuliert sind) ist der *Fragebogen zur Arbeitsanalyse* (FAA, Frieling & Hoyos, 1978). Die in dem Bogen enthaltenen Items oder Arbeitselemente lassen sich den Bereichen Informationsaufnahme und -verarbeitung, Arbeitsausführung, arbeitsrelevante Beziehungen sowie Umgebungseinflüsse und besondere Arbeitsbedingungen zuordnen und werden von Arbeitsanalytikern zumeist nach „Häufigkeit“, „Wichtigkeit“ und „Zeitdauer“ eingestuft. Der Fragebogen ist zur Analyse vielfältiger Berufstätigkeiten geeignet, jedoch können andere, vor allem komplexere und abstraktere Aufgaben mit diesem Bogen nicht vollständig beschrieben werden.

Ein im Vergleich dazu wenig standardisiertes, jedoch häufig verwandtes Verfahren zur verhaltensbezogenen Anforderungsanalyse ist die Critical Incident Technique (CIT, Flanagan, 1954). Die CIT zielt mit direkten Fragen darauf ab, erfolgskritisches Arbeitsverhalten zu identifizieren. Diese Technik soll im folgenden Abschnitt genauer erläutert werden.

Ein weiteres Verfahren der verhaltensbezogenen Anforderungsanalyse ist die sogenannte Delphi-Methode (Franke & Zerres, 1988). Sie basiert auf einer schriftlichen Expertenbefragung zu einem bestimmten Tätigkeitsbereich. Die Experten (Arbeitsplatz-inhaber, Vorgesetzte) wurden dazu befragt, welche Verhaltensweisen (und z.T. auch grundsätzliche Merkmale und Eigenschaften) eine Person für verschiedene Aspekte ihrer Tätigkeit an den Tag legen muss, um in diesem Tätigkeitsbereich ihre Aufgaben erfolgreich zu bewältigen. Die aus dieser Phase gewonnenen Informationen aller teilnehmenden Experten werden kategorisiert und allen Teilnehmern zugänglich gemacht. In einer zweiten

Phase können die Experten die gefundenen Merkmale ergänzen sowie diesen Merkmalen beobachtbares Verhalten zuordnen. Diese Merkmale und ihnen zugeordnete Verhaltensweisen werden wiederum geordnet und an die Experten versandt, so dass diese in der dritten Phase der Befragung die gesammelten Verhaltensweisen nach deren Wichtigkeit beurteilen können.

Eigenschaftsebene

Die dritte Ebene, auf welcher eine Anforderungsanalyse stattfinden kann, bezieht sich auf die Formulierung von Eigenschaften, die eine zentrale Rolle für die erfolgreiche Ausführung einer Tätigkeit spielen. Eine eigenschaftsbezogene Anforderungsanalyse beschäftigt sich also mit den Fähigkeiten, Fertigkeiten, Persönlichkeitseigenschaften und Interessen, die einen Einfluss auf die Arbeitsleistung sowie die -zufriedenheit nehmen. Charakteristisch für diesen Bereich sind auf der einen Seite *psychologische Testverfahren* zur Erfassung psychologischer Konstrukte, auf der anderen Seite die Entwicklung von *Eigenschaftslisten* zur Festlegung der Leistungskriterien. Ein Beispiel für eine solche Liste mit Eigenschaften sind die Ability Requirement Scales von Fleishman und Quaintance (1984), die kognitive, psychomotorische und physische Fähigkeiten umfassen. Ein weiteres Verfahren stellt die Synthetisierung eigenschaftsbezogener Anforderungsprofile auf der Basis einer Arbeitsanalyse dar. So kann der *Fragebogen zur Arbeitsanalyse* (FAA, Frieling & Hoyos, 1978) dazu verwendet werden, um eigenschaftsbezogene Anforderungprofile zu generieren. Hierfür wird für jedes der relevanten Arbeitselemente die Bedeutung jeder einzelnen Eigenschaft aus einer Liste eingeschätzt (vgl. Schuler, 2001).

Eine weitere Möglichkeit, eine eigenschaftsbezogene Anforderungsanalyse durchzuführen, ist die Orientierung am Fünf-Faktoren-Modell der Persönlichkeit (Raymark, Schmit & Guion, 1997): Hierbei wird die Erforderlichkeit einzelner Facettenmerkmale der jeweiligen Faktoren für den beruflichen Erfolg durch Experten eingeschätzt.

Einige methodische Probleme

Ein grundsätzliches Problem der Arbeits- und Anforderungsanalyse ist die Tatsache, dass die Bestimmung der Gütekriterien Objektivität, Reliabilität und Validität für diese Verfahren nur selten und unvollständig erfolgt. Wünschenswert hingegen wäre eine Ermittlung dieser Kriterien in analoger Weise zu den eignungsdiagnostischen Verfahren (vgl. Schuler, 2001). Ein weiteres methodisches Problem zeigt Landy (1993), indem er demonstrierte, dass Ergebnisse von Arbeitsanalysen auch von den Persönlichkeitsmerkmalen der Respondenten abhängen, die Effektstärke für das Geschlecht z.B. beträgt immerhin eine halbe Standardabweichung.

Trotz dieser methodischen Probleme soll an dieser Stelle erwähnt sein, das die Durchführung von Arbeits- und Anforderungsanalysen nicht nur als hilfreiche Grundlage in personalpsychologischen Belangen dient, sondern auch das Zustandekommen eines

Personalauswahlverfahrens oder einer Leistungsbeurteilung nachvollziehbar macht. Damit können solche Analysen, wie eingangs bereits erwähnt, als Maßnahmen der Qualitätssicherung angesehen werden (vgl. Schuler, 2001).

Die Critical Incident Technique

Die *Critical Incident Technique* (Flanagan, 1954) ist ein Verfahren zur verhaltensbezogenen Analyse von Anforderungen, das mit Hilfe gezielter Fragen erfolgskritisches Arbeitsverhalten zu identifizieren beabsichtigt, indem die Aufmerksamkeit auf das Verhalten besonders effektiver bzw. ineffektiver Stelleninhaber gelegt wird. Sie ist ein Verfahren, mit Hilfe dessen man wichtige Informationen bezüglich des Verhaltens in bestimmten Situationen sammeln kann. Mit *‚incident'*, auf deutsch am besten mit ‚Ereignis' zu übersetzen, ist in diesem Zusammenhang jedes beobachtbare menschliche Verhalten gemeint, das soweit in sich selbst vollständig ist, dass es Schlüsse auf die Person, die das Verhalten durchführt, zulässt. *‚Critical'* bedeutet im Zusammenhang mit diesem Verfahren, dass die beobachteten Verhaltensweisen eine entscheidende Rolle in Bezug auf Erfolg und Misserfolg spielen. Die Voraussetzung dafür ist, dass das ‚Ereignis' in einer Situation auftritt, in welcher der Zweck dieser Handlung offensichtlich ist und die Konsequenzen dieses Verhaltens klar sind.

Eine der ersten Studien, die mit der Critical Incident Technique durchgeführt wurden, fand im Rahmen des „Aviation Psychology Programm" der United States Army Air Forces während des Zweiten Weltkriegens statt und diente der Analyse von Ursachen für Versagen in der Fliegerausbildung, um auf deren Basis Auswahltests zu entwickeln.

Die Durchführung einer Anforderungsanalyse mit der Critical Incident Technique umfasst fünf Abschnitte, die im Folgenden näher erläutert werden sollen.

Die Formulierung eines grundlegenden Ziels

Die Critical Incident Methode hat zum Ziel, Verhaltensweisen von Arbeitsplatzinhabern bzw. Studenten im Hinblick auf Effektivität zu beurteilen und zu sammeln. Allerdings ist es offensichtlich unmöglich zu berichten, dass eine Person bei der Ausführung einer Aufgabe entweder besonders effektives oder besonders ineffektives Verhalten gezeigt hat, sofern unbekannt ist, welches Ziel diese Person verfolgen soll. Bspw. mag das Verhalten eines Vorgesetzten, der einem Arbeiter einen halben Tag frei gibt, um an einer Erholungsaktivität teilzunehmen, als sehr effektiv eingeschätzt werden, wenn es das Hauptziel dieses Vorgesetzten ist, gut mit seinen Mitarbeitern zurecht zu kommen. Dasselbe Verhalten jedoch mag als ineffektiv eingeschätzt werden, wenn das vordergründige, grundlegende Ziel in der unmittelbaren Produktion von Gütern oder im Service liegt.

Für die meisten Situationen gibt es kein Hauptziel, welches das einzig korrekte ist. Unterschiedliche Personengruppen definieren unterschiedliche grundlegende Ziele der gleichen Tätigkeit. So haben Studenten, Dozenten und potentielle spätere Arbeitgeber eine unterschiedliche Vorstellung davon, worin das hauptsächliche Ziel eines Studenten eines bestimmten Faches liegt. Mögliche primäre Ziele könnten z.B. ein guter Notendurchschnitt, eine kurze Studiendauer, Zufriedenheit mit dem Studium oder ein hohes Engagement in mit dem Studium verwandten, praktischen Bereichen (Therapieeinrichtungen, Beratungszentren, Forschungsinstitute, …). Um eine Anforderungsanalyse mittels der Critical Incident Methode durchführen zu können, sollten einerseits die teilnehmenden Experten eine ähnliche Vorstellung von dem Hauptziel der betreffenden Tätigkeit haben, andererseits sollte dieses Ziel aber auch an die möglichen Verwender einer solchen Analyse deutlich kommuniziert werden. Auf Basis einer Anforderungsanalyse, die erfolgskritische Verhaltensweisen eines zufriedenen Studenten ermittelt, lässt sich weniger gut ein Eignungstest für die schnellsten Studenten entwickeln. Flanagan (1954) schlägt vor, im Vorfeld einer Anforderungsanalyse ein generelles Ziel einer Tätigkeit, mit welchem die meisten Menschen übereinstimmen würden, durch Experten formulieren zu lassen.

Planung und Spezifizierung

Um die wirklich zentralen Aspekte bei der funktionalen Beschreibung eines Verhaltens zu ermitteln, muss der Beobachter präzise Instruktionen erhalten. So hat es sich nach Flanagan (1954) als sinnvoll erwiesen, die Beobachter bei der Sammlung kritischer Verhaltensweisen nur *außerordentlich* effektive oder ineffektive Verhaltensweisen zu sammeln. Extreme Ereignisse können nämlich genauer identifiziert werden als Verhalten, das nahe am Durchschnitt liegt.

Bei der Planung, welche Personen man als Beobachter einsetzen möchte, sollte ein Kriterium die Erfahrung im betreffenden Tätigkeitsbereich sein. Als beste Beobachter haben sich diejenigen Personen erwiesen, die in ihrer eigenen Beschäftigung die Aufgabe haben, dass Arbeitsverhalten von Personen im gefragten Tätigkeitsbereich zu beobachten und zu beurteilen (Vorgesetzte, Supervisoren).

Auch in Bezug auf die zu beobachtende Situation, deren Relevanz in Bezug auf das Hauptziel der Tätigkeit, sowie das Ausmaß des Effektes einer Verhaltensweise auf diese Ziel sollten für die Beobachter genau spezifiziert sein.

Bezüglich der Situation sollten dem Beobachter Informationen gegeben werden über den Ort, die beteiligten Personen, die Bedingungen und die Aktivitäten. Ein Beispiel für eine solche Information ist Folgendes: „Studenten sollen beobachtet werden während eines Seminars, in welchem sie anderen Studenten ein Referat halten.“

Damit die Beobachter wissen, wann sie eine der oben beschriebenen zu beobachtenden Verhaltensweisen in einer bestimmten Situation als relevant in Bezug auf das primäre Ziel

einzustufen haben, sollten sie z.B. Informationen darüber erhalten, ob eine Verhaltensweise direkten Einfluss auf das Ziel haben sollte oder ob auch ein indirekter Einfluss als ‚kritisch' bezeichnet wird und ob der Effekt unmittelbar oder langfristig sein soll.

Bezüglich des Ausmaßes des Effektes, das ein Verhalten auf das primäre Ziel einer Tätigkeit haben sollte, um als ‚critical incident' bezeichnet werden zu können, hat es sich nach Flanagan (1954) als hilfreich erwiesen, den Beobachtern den Hinweis zu geben, dass eine Verhaltensweise dann als kritisch zu bezeichnen ist, wenn sie einen „bedeutsamen" Beitrag (sowohl positiv als auch negativ) zu dem Ziel der Tätigkeit leistet. In einigen Fällen ist es möglich, ein quantitatives Kriterium zu nennen, z.B. eine Mindestnote, die ein Student erreicht haben muss, um über sein Verhalten zu berichten zu können.

Eine große Wichtigkeit kommt auch dem Training der Beobachter zu, in welchem sie nochmals auf das primäre Ziel der betreffenden Tätigkeit sowie auf die oben genannten Spezifizierungen bezüglich Situation, Relevanz und Ausmaß des Effektes hingewiesen werden.

Datenerhebung

Für die Phase, in der die Beobachter die ‚kritischen Verhaltensweisen' sammeln, hat es sich als günstig erwiesen, die zu beobachtenden Ereignisse möglichst zeitnah aufzuzeichnen. Flanagan (1954) berichtet, dass Beobachter, die kritische Verhaltensweisen erst nach einer 2-wöchigen Periode aufzuzeichnen hatten, 80% weniger Ereignisse berichteten als diejenigen Beobachter, die zwei Wochen lang jeden Tag beobachtete kritische Verhaltensweisen aufzeichneten. Meist bekommt man jedoch keine vollständige Information über eine Tätigkeit, wenn man ausschließlich kürzlich aufgetretene Ereignisse aufzeichnet. Ein Anhaltspunkt, dass auch weiter zurückliegende Ereignisse, relativ genau berichtet wurden, ist eine umfang- und detailreiche Beschreibung, vage Berichte hingegen lassen eher darauf schließen, dass das Ereignis nicht gut erinnert ist.

Im Folgenden soll kurz auf vier verschiedene Möglichkeiten der Datenerhebung, Einzel- und Gruppeninterview, Fragebogenerhebung, sowie schriftliche Aufzeichnungen, eingegangen werden.

Flanagan (1954) bezeichnet die Datenerhebung mittels *Einzelinterviews* durch geschulte Interviewer, die Beobachtern genau erklären können, welche Informationen erwünscht sind und ihnen alle nötigen Details vermitteln können, als die befriedigendste. Bei dieser Vorgehensweise ist es wichtig, den Beobachter zunächst über den Auftraggeber sowie den Zweck der Studie zu informieren und ihm des weiteren darüber aufzuklären, aus welchen Gründen er für diese Aufgabe besonders qualifiziert ist und daher für die Erhebung ausgewählt wurde. Auch ist es sinnvoll, dem Beobachter die Daten die Handhabung der Daten zu erklären, so dass dieser sich von der Anonymität seiner Angaben überzeugen kann und nicht wichtige Informationen zurückhält, aus Angst, jemandem zu schaden. Bei der

Formulierung der Aufgabenstellung, nämlich erfolgskritisches Arbeitsverhalten in definierten Situationen zu sammeln, ist besondere Vorsicht geboten. Flanagan (1954) berichtet von einer Studie, in welcher die Frage gestellt wurde, wie sich ein Arbeiter *verhielt (‚behaved')*, so dass es zu Produktionssteigerungen kam, oder aber was er dazu *tat (‚did')*. Anscheinend ließ die erste Frage die Beobachter annehmen, Persönlichkeit und Einstellungen würden untersucht, während die zweite Frage zu einer umfassenderen Beschreibung von kritischen Verhaltensweisen führte. Die Hauptaufgabe des Interviewers sollte es also sein, eindeutig zu vermitteln, welche Art von Verhalten beobachtet und berichtet werden sollte, im weiteren Verlauf des Gespräches sollte der Interviewer nur noch dann eingreifen, wenn der Beobachter die Aufgabenstellung nicht richtig verstanden hat. In Abb. 2 ist ein Leitfaden für Interviewer zur Erhebung von kritischen Ereignissen dargestellt.

"Think of the last time you saw one of your subordinates do something that was very helpful to your group in meeting their production schedule. "

(Pause till he indicates he has such an incident in mind.) "Did his action result in increase in production of as much as one per cent for that day?"

(If the answer is "no", say) "I wonder if you could think of the last time that someone did something that did have this much of an effect in increasing production."

"What were the general circumstances leading up to this incident?"

_____

"Tell me exactly what this person did that was so helpful at that time."

_____

"Why was this so helpful in getting your group's job done?"

_____

"When did this incident happen?"

_____

"What was this person's job?"

_____

"How long has he been on this job?"

_____

"How old is he?"

_____

Abb. 2. Beispiel eines Interviewleitfadens zur Sammlung kritischer Ereignisse (Flanagan, 1954)

Gruppeninterviews wurden eingesetzt, um Zeit und Personal zu sparen. Die Vorteile des Einzelinterviews wie persönlicher Kontakt und Erklärung, die ständige Verfügbarkeit des Interviewers um Fragen zu beantworten, bleiben dabei erhalten. Zumeist wird so vorgegangen, dass der Interviewer eine einführende Erklärung und die Aufgabenstellung wie beim Einzelinterview gibt, anschließend werden die Beobachter gebeten, die kritischen Ereignisse niederzuschreiben anhand von dem auf Abb.2 dargestellten Leitfaden ähnlichen Vorgaben. Der Interviewer kann die Antworten auf die erste Frage anschauen, um auf diese Weise sicherzugehen, dass die Aufgabe richtig verstanden wurde.

Fragebögen sind vor allem dann einzusetzen, wenn die Gruppe der Beobachter größer wird. Bei zugesendeten Fragebögen ist es wichtig, die Beobachter zu motivieren, die Instruktionen sorgfältig zu lesen und zu befolgen.

Eine weitere Möglichkeit ist die der *schriftlichen Aufzeichnungen*, eine Variante davon ist die schriftliche Aufzeichnung einer kritischen Verhaltensweise durch den Beobachter in dem Moment, in welchem sie auftritt.

Bezüglich der Anzahl von kritischen Ereignissen, die erforderlich ist, um die Anforderungen einer Tätigkeit umfassend zu umschreiben, lässt sich keine eindeutige Angabe machen. Als Anhaltspunkt gibt Flanagan (1954) die Auskunft, dass eine recht einfache Tätigkeit mit 50-100 kritischen Verhaltensweisen beschrieben werden kann, während es zur Beschreibung komplexer Tätigkeiten einer Stichprobe von mehreren tausend kritischen Verhaltensweisen bedarf.

Datenanalyse

Nachdem die Daten gesammelt worden sind, müssen sie zusammengefasst und beschrieben werden, so dass sie für unterschiedliche praktische Zwecke effektiv genutzt werden können. Dabei sollte man sich über die Art der Klassifikation, die man verwenden möchte, über die Formulierung der Kategorien sowie über den Grad der Detailliertheit der resultierenden Kategorien im Klaren sein.

Je nachdem zu welchem Zweck eine Critical Incident Analyse durchgeführt wird (Auswahl, Training, Leistungsbeurteilung), sind andere *Arten von Klassifikationen* zu verwenden. So sollten für Auswahlzwecke solche Kategorien bestimmt werden, die sich leicht auf die psychologischen Traits, die für den Auswahlprozess verwendet werden, beziehen lassen. Gleichermaßen sollten die Kategorien für Trainingszwecke so formuliert sein, dass sie leicht auf Trainingskurse und –ziele übertragbar sind. Ähnliches gilt auch für die Leistungsbeurteilung, hier sollte eine größere Aufmerksamkeit auf die Komponenten der Arbeit, wie sie im Moment ausgeführt wird, gelegt werden.

Für die *Formulierung der Kategorien* schlägt Flanagan (1954) folgendes Vorgehen vor: Eine relativ kleine Menge der gesammelten kritischen Ereignisse wird herausgegriffen und die darin enthaltenen ‚kritischen Ereignisse' werden in Gruppen mit ähnlichen

Verhaltensweisen sortiert, die einen Bezug haben zu der Art der Klassifikation, die vorher gewählt wurde. Anschließend werden mit Hilfe von Experten die gefundenen Kategorien bestätigt, bzw. modifiziert und definiert. Danach können die übrigen Verhaltensweisen zugeordnet werden. Wenn nötig, können größere Kategorien nochmals in Untergruppen differenziert werden. Schließlich sollte der *Grad der Detailliertheit bzw. Generalität* der Kategorien bestimmt werden. Hierbei sollten die Titel der einzelnen Anforderungskategorien logisch organisiert sein, die Titel allein sollten bereits ohne weitere Erklärung die Bedeutung der Kategorie übermitteln, die Kategorien sollten alle in etwa die gleiche Wichtigkeit besitzen, die Gesamtheit der Kategorien sollte alle häufig genannten kritischen Verhaltensweisen abdecken und die Kategorien sollten so benannt werden, dass die für weitere Forschung und Entwicklung möglichst leicht anwendbar und nützlich ist.

Interpretation und Darstellung

Es ist kaum möglich, bei der Erstellung einer Anforderungsanalyse für alle auftretenden praktischen Probleme eine ideale Lösung zu finden. Deshalb bedürfen die Ergebnisse einer angemessenen Interpretation, um sie sinnvoll weiterverwenden zu können. So müssen die einzelnen Schritte, wie die Bestimmung des primären Ziels der betreffenden Tätigkeit, die Planung und Spezifizierung, die Datensammlung und –analyse klar kommuniziert werden, so dass potentielle Anwender sich im Klaren sein können, ob und wie sie mit den Ergebnissen umzugehen haben. Auch sollte deutlich herausgestellt werden, wenn die Personen, an welchen die Beobachtungen angestellt wurden, nicht repräsentativ sind für die Gesamtheit der Personen, die die betreffende Tätigkeit ausführen. Während die Einschränkungen der Verwendung der Anforderungsanalyse deutlich zu berichten ist, sollte der Wert der Ergebnisse für weitere Untersuchungen und Entwicklungen ebenfalls hervorgehoben werden.

Verfahrensbeispiel Anforderungsanalyse an der RWTH Aachen

Als praktisches Beispiel soll an dieser Stelle eine Anforderungsanalyse vorgestellt werden, die als Grundlage für die Weiterentwicklung eines Internet – Selbsteinschätzungs – Test zur Studienneigung für die Fächer Elektrotechnik, technische Informatik sowie Informatik diente (Zimmerhofer, 2003).

Zimmerhofer führte zunächst eine *Dokumentenanalyse*, welche zu den unstandardisierten Verfahren der Anforderungsanalyse zählt, durch. Auf diese Weise konnten Informationen über die Ziele der jeweiligen Studiengänge, den Studienaufbau, die involvierten Fächer, die Art der Veranstaltungen sowie über die zu erbringenden Leistungsnachweise gesammelt werden.

Als nächster Schritt wurde eine *Expertenbefragung* durchgeführt, Professoren, Dozenten, wissenschaftliche Mitarbeiter und Studenten stellten dabei die Experten dar, die zunächst in einem halbstandardisierten Interview direkt nach den zentralen Anforderungen der Studiengänge in den Bereichen ‚Verhaltensanforderungen', ‚Eigenschaftsanforderungen', ‚Ergebnisanforderungen' und Qualifikationsanforderungen befragt.

Ferner wurde die oben besprochene *Critical Incident Technique* (Flanagan, 1954) zur Unterscheidung erfolgreicher und erfolgloser Studenten in typischen Ausgangssituationen eingesetzt, um Ursachen für Erfolg und Misserfolg im Studium zu ergründen.

Ein weiterer Themenkomplex in den Interviews stellte die *Erfahrung mit Studienabbrechern* und typischen Hürden für Studierende v.a. im Grundstudium dar.

Anschließend wurde den Experten die Möglichkeit gegeben, eigene *Anregungen zur Gestaltung der Aufgaben* für den Selbsteinschätzungstest zu nennen.

Bei den Interviews wurde darauf geachtet, dass die Experten die übergeordneten Anforderungen wie z.B. „logisches Denken" durch Beispiele, Nennung konkreter Verhaltensweisen, und Definitionen mit Inhalt füllen.

Zusätzliche Gespräche wurden mit Vertretern der Studienberatung, des Arbeitsamtes sowie Experten im Bereich Studienneigung und –Beratung geführt.

Anschließend wurden die gesammelten Merkmale nach inhaltlichen Gesichtspunkten gruppiert und mit Überschriften versehen, um einen besseren Überblick zu schaffen sowie die strukturellen Beziehungen der Anforderungen zu verdeutlichen.

Abschließend wurden die identifizierten Merkmalsgruppen nach ihrer Bedeutung für den Studienerfolg gewichtet.

Eine Darstellung der Anforderungen erfolgte in Anlehnung an Blums informationstheoretisch orientierte Grobstruktur (1980), die zwischen Informationsaufnahme, -Verarbeitung, -Speicherung, -Wiedergabe und sonstigen Merkmalen unterscheidet, wobei die Informationsverarbeitung den umfangreichsten und am stärksten differenzierten Merkmalsbereich darstellt. Das resultierende Anforderungsprofil ist in Tabelle 1 dargestellt.

Tabelle 1: Anforderungsprofil der Elektrotechnik/Technische Informatik und der Informatik (vgl. Zimmerhofer, 2003)

Anforderungsdimension	Zusammenfassende Ergebnisse der Anforderungsanalyse für E-Technik/Technische Informatik	Zusammenfassende Ergebnisse der Anforderungsanalyse für Informatik
INFORMATIONSAUFNAHME		
1. Beobachtungs- und Auffassungsvermögen	Schnelles Einarbeiten in neue Themen Problemorientierung: Wahrnehmung von Problemen	
INFORMATIONSVERARBEITUNG		
2. Analysieren	Einen Sachverhalt verständnismäßig genau erfassen Logische Zusammenhänge erkennen Komplexe Sachverhalte und Zusammenhänge verstehen Besondere Eigenschaften eines Sachverhalts extrahieren …	
3. Systematisieren	Wichtiges von Unwichtigem trennen Essenz auf Informationen ziehen Strukturen in Daten erkennen …	
4. Formalisieren	Modellbildung: ein nichtformalen Modell in ein formales Modell überführen Konkrete Sachverhalten in formale Begriffe fassen Konkrete Sachverhalte mathematisieren	
5. Abstrahieren und induktives Denken	Übergeordnete Prinzipien erkennen Aus Beobachtungen Regeln ableiten Denken in Modellen	
6. Deduktives Denken	Neues Wissen aus gegebenem ableiten Wenn-Dann-Zusammenhänge verstehen Aus abstrakter Gegebenheit schlussfolgern ..	
7. Konkretisieren abstrakter Sachverhalte (Anwendungsbezogenheit)	Modelle auf konkreten Fall anwenden Praxisorientiert denken Konsequenzen einer Aktion antizipieren …	
8. Kombinieren und integrieren	Verschiedene Theorien verknüpfen und anwenden Informationen aus verschiedenen Bereichen verarbeiten und miteinander in Beziehung setzen	
9. Planen und organisieren	Überblick behalten Arbeitsökonomie und Selbstorganisation beim Lernen Inhalte strukturieren …	
10. Kreativität und Flexibilität	Kreativität beim Problemlösen (Ideenreichtum, ungewöhnliche Ideen)	

		Fantasie/Vorstellungsvermögen Denkschemata über Bord werfen …	
11.	*Operieren mit Formeln, Regeln, Größen und Größenordnungen*	Prinzipielles Verständnis für Mathematik Fähigkeit, Größen abzuschätzen Umgang mit Formeln	
12.	*räumliches Vorstellungsvermögen*	Räumliches Denken	
13.	*Fähigkeit, sich technisch-naturwissenschaftliche Abläufe und Zustände vorzustellen*	Ablauf von Vorgängen aufgrund von verbalem oder visuellem Material nachvollziehen Abläufe verbal oder grafisch dokumentieren	
14.	*Beurteilen und Kritisieren*	Eigene Leistungsfähigkeit treffend einschätzen Machbarkeit von Lösungen beurteilen	
INFORMATIONSSPEICHERUNG			
15.	*Gedächtnis, Merkfähigkeit*	Große Gedächtniskapazität	
16.	*Kenntnisse*	Grundkenntnisse in Mathematik Englischkenntnisse	
17.	*Praktische Erfahrungen*	Umgang mit dem PC Fingerfertigkeit	

Auf ein völlig einheitliches Anforderungsprofil für Elektrotechnik/Technische Informatik und Informatik wurde verzichtet, da sich in einigen Bereichen („räumliches Vorstellungsvermögen“, „Fähigkeit, sich technisch-naturwissenschaftliche Abläufe vorzustellen“) Unterschiede ergaben. Die meisten Anforderungsdimensionen waren jedoch für beide Bereiche identisch.

Literatur

Amelang; M. (1997). Differentielle Aspekte der Hochschulzulassung: Probleme, Befunde; Lösungen. In Th. Hermann (Hrsg.), *Hochschulentwicklung – Aufgaben und Chancen* (S. 88-105). Heidelberg: Asanger.

Blum, F. (1980). *Studieneignung für die Ingenieurswissenschaften.* München: Minerva Publikation.

Braun, O. & Maaßen, J. (2000). Anforderungsanalyse für Existenzgründer. In Müller, G. F. (Hrsg.), Existenzgruendung und unternehmerisches Handeln. Forschung und Förderung (S. 37-52). Landau: Verlag Empirische Paedagogik.

Flanagan, J. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327-358.

Fleishman, E.A. & Quaintance, M.K. (1994). *Taxonomies of human performance.* Orlando, FL : Academic Press.

Franke, R. & Zerres, M.P. (1988). *Planungstechniken: Instrumente für zukunftsorientierte Unternehmensführung*. Aschaffenburg: FAZ-Verlag.

Frieling, E. & Hoyos, C.G. (1978). *Fragebogen zur Arbeitsanalyse (FAA): Deutsche Bearbeitung des Position Analysis Questionnaire (PAQ)*. Bern: Huber.

Jeserich, W. (1981*). Mitarbeiter auswählen und fördern: Assessment-Center-Verfahren*. München: Hanser.

Landy, F.J. (1993). Job analysis and job evaluation: The respondent's perspective. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 75-90). Hillsdale, NJ: Erlbaum.

Raymark, P.H., Schmit, M.J. & Guion, R.M. (1997). Identifying potentially useful personality constructs for employee selection. *Personnel Psychology, 50,* 723-736.

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.

Schuler, H. (1989). Some advantages and problems of job analysis. In M. Smith & I.T. Robertsond (Eds.), *Advances in Selection and Assessment* (pp. 31-42). New York: Wiley.

Schuler, H. & Funke, U. (1995). Diagnose beruflicher Eignung und Leistung. In H. Schuler (Hrsg.), *Lehrbuch Organisationspsychologie* (2. Aufl., S. 235-283). Bern: Huber.

Schuler, H. (2001). Arbeits- und Anforderungsanalyse. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 43-61). Göttingen: Hogrefe.

Viek, P., Heller, K., Becker, U. &Schober, B. (1997). *Explorationsstudie zur Begabtenförderung im Tertiärbereich.* München: Abschlussbericht an das BMBF.

Zimmerhofer, A. (2003). *Neukonstruktion und erste Erprobung eines webbasierten SelfAssessments zur Feststellung der Studienneigung für die Fächer Elektrotechnik, Technische Informatik sowie Informatik an der RWTH Aachen,* Unveröffentlichte Diplomarbeit, Psychologisches Institut, RWTH Aachen.

Psychologische Leistungstests & Schulnoten

Helge Sickmann

Einleitung

Bei einer Studierendenauswahl stellt sich wie bei jeder Bewerberselektion die Frage nach den Kriterien des Erfolgs, nach den Prädiktoren eben dieser Kriterien, sowie der Größe des Zusammenhangs zwischen beiden. Gegenstand dieser Arbeit werden die Prädiktoren von Studienerfolg sein.

In diesem Rahmen wird die prognostische Validität von Abiturnoten und Leistungstest erörtert werden. Auf die Rolle weiterer Prädiktoren, etwa von Persönlichkeitsmerkmalen, anderer allgemeiner Kompetenzen oder sozioökonomischer Faktoren wird verzichtet, da diese Prädiktoren für die Hochschulzulassung in Deutschland nicht vorgesehen sind. Dennoch sei an dieser Stelle darauf hingewiesen, dass verschiedene Studien die Validität gerade solcher Prädiktoren zeigen.

Auch wird auf die Bedeutung von Persönlichkeitstests und Interviews als diagnostische Methoden in dieser Arbeit nicht eingegangen, Interviews sind Inhalt einer anderen Hausarbeit.

Allgemein beruht die Verwendung diagnostischer Verfahren beim Hochschulzugang auf folgenden Annahmen (Deidesheimer Kreis, 1997):

- Es gibt eine Reihe von Merkmalen, die in je unterschiedlicher Weise für den Erfolg in verschiedenen Studiengängen wichtig sind.
- Diese Eignungsmerkmale sind bei Studienbewerbern unterschiedlich ausgeprägt, der jeweilige Ausprägungsgrad lässt sich mit Hilfe entsprechender diagnostischer Instrumente messen bzw. abschätzen.
- Die Eignungsmerkmale sind relativ überdauernd und erlauben mithin eine längerfristige Prognose.
- Fähigkeiten sind Kernbestandteile der Eignungsmerkmale. Sie können sich auf verschiedene Weise entwickelt haben bzw. erworben worden sein. Rückschlüsse auf das vorliegen solcher Fähigkeiten können auch aus bereits erbrachten schulischen und außerschulischen Leistungen gezogen werden.

Abiturnote:

Die Leistung im Abitur ist in fast allen Staaten, die ein Hochschulwesen besitzen, ein Kriterium für die Hochschulzulassung.

Schulnoten dienen der Leistungsbeurteilung in einem begrenzten Anforderungsgebiet. In Deutschland wurde bisher, in überlasteten Studienfächern, die Studierendenauswahl fast ausschließlich zentral durch die Zentralstelle für die Vergabe von Studienplätzen (ZVS) getroffen. Hierbei zählten, von einigen sozialen Sonderregelungen abgesehen, nur die durchschnittliche Abiturnote sowie die Wartesemester. Die Abiturnote als Prädiktor wird verwendet, da davon ausgegangen wird, dass Schüler mit guten Noten auch erfolgreich im Studium sind.

Schulnoten an sich haben jedoch nicht die Aufgabe spätere Leistungen zu prognostizieren. Die Bewertung des Schülers durch den Lehrer sollte sich nur auf die direkt beobachteten Leistungen und nicht auf Erwartungen im Hinblick auf späteren Ausbildungs- oder Berufserfolg beziehen. Hinter diesen Leistungen stehen jedoch relativ stabile Merkmale wie motivationale Faktoren, Anstrengung, Fähigkeitsfaktoren, konstitutionelle Faktoren und Persönlichkeitsmerkmalen. Deshalb ist anzunehmen, dass die Abiturnoten neben der inhaltlichen Validität auch prognostische Validität bezüglich späterer Leistungen besitzen. Dies gilt besonders bei Leistungen die strukturell ähnlich zu schulischen Anforderungen sind (vgl. Schuler 1998). Für das Erreichen von Studienleistungen ist diese strukturelle Ähnlichkeit gegeben, da für das erfolgreiche Abschließen eines Studiums viele Faktoren notwendig sind, die auch einen erfolgreichen Schulabschluss bedingen. Daher sollte eine recht hohe prognostische Validität bezüglich des späteren Studienerfolges zu erwarten sein.

Schulische Leistungen können in verschiedener Art und Weise bei der Entscheidung über die Hochschulzulassung eingesetzt werden. Als Auswahlkriterien können, einzeln oder in verschiedenen Kombinationen, der Notendurchschnitt im Abschlusszeugnis, einzelne Fachnoten, belegte Kurse bzw. Fächerwahl, die Rangposition in der Abschlussklasse (diese Information liegt in Deutschland jedoch nicht vor), oder Schulgutachten bzw. Empfehlungsschreiben verwendet werden (Deidesheimer Kreis 1997).

Im Folgenden werden daher zunächst Vorteile sowie Kritikpunkte des Prädiktors Schulleistung dargestellt, wobei auf die Abiturdurchschnittsnote und auf Einzelfachnoten eingegangen wird.

Abiturdurchschnittsnote:

Vorteile der Abiturdurchschnittsnote gegenüber Einzelfachnoten sind (vgl. Rindermann und Oubaid, 1999):

1. Die Durchschnittsnote ist durch ein höheres Aggregationsniveau messgenauer. Spezifische Einflüsse einzelner Lehrerurteile und Prüfungsergebnisse werden ausgemittelt (Baron-Boldt et al., 1988)
2. Die Abiturdurchschnittsnote ist bei Bewerbern mit Abitur immer vorhanden. Fachnoten hingegen liegen aufgrund der Abwählbarkeit einzelner Fächer nicht immer vor. Zudem stellt die Vergleichbarkeit von Leistungs- und Grundkursen ein Problem dar.
3. Die Abiturdurchschnittsnote spiegelt eher als Einzelfachnoten Allgemeinbildung und allgemeine kognitive, als auch nichtkognitive Kompetenzen und motivationale Einstellungen wider, welche sowohl für einen erfolgreichen Schulbesuch, wie auch für ein erfolgreiches Studium notwendig sind. Dazu zählen Arbeitshaltung, Motivation, Fleiß, Anpassung, Arbeitsmanagement u.a.. Diese generellen Leistungsvoraussetzungen sind für alle Studienfächer unabdingbar.

Kritikpunkte an der Abiturnote gibt es zahlreiche. Beispielsweise führen die in den einzelnen Bundesländern verschieden gestalteten gymnasialen Oberstufen dazu, dass die Vergleichbarkeit von Abiturnoten nur schwer möglich ist. Problematisch ist zudem die an verschiedenen Schulen unterschiedlich geregelte Abwählbarkeit einzelner Fächer sowie die verschiedenen Kurskombinationen. Immer wieder wird auch auf die unterschiedliche Schwierigkeit des Abiturs in den Bundesländern, aber auch an einzelnen Schulen, hingewiesen. Aus diesen Gründen ist es zweifelhaft, ob die Abiturnote eine allgemeingültige Aussagekraft bezüglich der Studierfähigkeit für alle Studienfächer besitzt (Althoff, 1986).

Einzelfachnoten:

Einzelfachnoten könnten als Prädiktoren verwendet werden, da angenommen wird, dass zwischen dem Erfolg in inhaltlich korrespondierenden und Schul- und Studienfächern eine engere Beziehung besteht als zwischen Abiturdurchschnittsnoten und dem fachspezifischen Studienerfolg. So erscheint beispielsweise die Sportnote augenschein- valider für die Prognose der Studierfähigkeit der Sportwissenschaften als etwa die Schulnote in Geschichte.

Empirische Befunde:

Die bisherigen Studien zu Studienerfolg zeigen, dass die Abiturnote eine der besten und ökonomischsten Prädiktoren sowohl für einen schnellen Studienabschluss, einen guten Studienabschluss und eine geringe Abbruchsquote ist.

Die Abiturnote als Prädiktor für den allgemeinen Studienerfolg:

Verschiedene Studien zur prognostischen Validität der Abiturnote zeigten recht verschiedene Ergebnisse. Trost und Bickel (1979) stellten 52 Einzelvergleiche zwischen Abiturdurchschnittsnote und Vorexamens- bzw. Examensdurchschnittsnote in verschiedenen Studienfächern zusammen. Die Korrelationskoeffizienten der einzelnen Studien lagen zwischen r= - .02 und r= .53, als Mittelwert wurde r=.35 errechnet.

Um verlässlichere Ergebnisse über die prognostische Validität von Abiturnoten zu erhalten führten Baron-Boldt, Schuler und Funke (1988) eine Metaanalyse nach der Methode der Validitätsgeneralisierung nach Hunter und Schmidt durch.

Diese Methode wurde speziell zur Integration von Korrelationskoeffizienten über Einzelstudien hinweg entwickelt. Durch diese Integration kommt es zu einer Summierung der Stichprobengrößen der Einzelstudien und somit zu einer höheren Teststärke. Ein weiterer entscheidender Vorteil ist die Möglichkeit, bei der Zusammenfassung der Validitäten eine Korrektur für verschiedene Artefakte durchzuführen. Bei der Methode der Validitätsgeneralisierung können Korrekturen von Stichprobenfehler, die durch die geringen Stichprobengrößen der Einzelstudien zustande kamen, vorgenommen werden. Des Weiteren sind Korrekturen für die Ungenauigkeit der Messungen von Prädiktoren und Kriterien und für die Messbereichseinschränkungen aufgrund der Auswahl vorgesehen. Die Korrekturen der Messbereichseinschränkungen wurden durchgeführt, da die Varianz der Abiturnoten in den Stichproben der Studierenden kleiner ist, als in der Bewerberpopulation. Der korrelative Zusammenhang zwischen Abiturnote und Studienerfolg unterschätzt folglich die prognostische Validität, dieses soll durch die Korrekturen ausgeglichen werden.

In dieser Metaanalyse wurden insgesamt 61 deutschsprachige Einzelstudien zusammengeführt. Es resultierte eine Stichprobengröße von insgesamt N= 30122 (median= 137, Streubreite: n=11 – n= 4688). Zur Prognose des Studienerfolgs durch die Abiturnote wurden 44 Studien verwendet (N=26867, median n=160, Streubreite n=12 – n= 4688).

Die Ergebnisse der Metaanalyse zeigten, dass die große Varianz in den Validitäts-koeffizienten der Einzelstudien nur zum Teil auf wahre Unterschiede basieren, zu einem größeren Teil jedoch durch Stichprobenfehler und andere Artefakte verursacht wurden.

Zu bedenken ist jedoch, dass durch die in der Metaanalyse durchgeführten Artefaktkorrekturen nicht sämtliche Ungewissheiten beseitigt werden konnten, da den Einzeluntersuchungen zum Teil verschiedene Prädiktoren, Kriterien, Zeiträume und sonstige Bedingungsgrößen zugrunde liegen.

Die mittlere gewichtete, noch unkorrigierte Validität lag bei r=.345, mit einer Varianz von s_r^2 = .0078. Nach einer Korrektur des Stichprobenfehlers, der Unreliabilität der Kriteriumsmessung, sowie der Einschränkung des Messbereichs, ergibt sich eine mittlere artefaktkorrigierte Validität von .456, mit einem Konfidenzintervall (p=0.05) von .317 bis .595 und einer Varianz von .005. Auf eine Korrektur für die Unreliabilität der Prädiktoren wurde verzichtet, da bei Verwendung der Abitursnoten zur Studienerfolgsprognose die Ungenauigkeit der Messung auch nicht ausgeschlossen werden kann.

Ähnliche Ergebnisse fanden Robins et al (2004) in einer Metaanalyse von 109 amerikanischen Einzelstudien. Als Prädiktor diente der High School GPA, als Kriterium die durchschnittlichen Noten (GPA) im College. Die Daten zeigten eine prognostische Validität von .413. Die prognostische Validität in dieser Metaanalyse wird jedoch unterschätzt, da keine Korrekturen für Messbereichseinschränkungen vorgenommen wurden. Dies kann das etwas geringere Ergebnis erklären.

Prognostische Validität bzgl. verschiedener Studiengänge:

In der Metaanalyse von Baron-Boldt, Schuler und Funke (1988) zeigte sich des Weiteren, dass der Prädiktor Abiturnote für verschiedene Studiengänge unterschiedlich valide ist.

	N	k	$\bar{\varrho}$	σ^2	%	Δ_{crit} ($\bar{\varrho}$)
Gesamtgruppe	26867	75	.456	.005	64	.317–.595
Moderatorvariable: Studienfach						
Psychologie	1187	10	.455	-.008	185	.455–.455
Medizin	4677	18	.448	.022	32	.157–.739
Geisteswissensch. und Lehrämter	1298	15	.460	-.009	181	.460–.460
Wirtschaftswiss.	1067	4	.557	.006	67	.405–.709
Jura	808	2	.377	.026	21	.061–.693
Mathematik und Naturwissenschaft	3682	17	.446	.003	78	.339–.553
Moderatorvariable: Veröffentlichungsform						
Veröffentl. Studien	6413	37	.431	.017	42	.175–.687
Unveröff. Studien	6863	31	.470	.005	70	.331–.609

N Größe der Gesamtstichprobe der jeweiligen Analyse
k Anzahl der unabhängigen Stichproben in der jeweiligen Analyse
$\bar{\varrho}$ Mittlere korrigierte Validität
σ^2 Korrigierte Varianz
% Prozentualer Anteil der Gesamtvarianz, der durch alle Artefakte erklärt wird
Δ_{crit} Konfidenzintervall mit p = 95 % um $\bar{\varrho}$

Abbildung 1: Validitäten für verschiedene Studienfächer in der Studie von Baron-Boldt, Schuler und Funke (1988, S. 82).

Wie Abb. 1 zeigt, ergaben sich die stärksten Abweichungen von der mittleren Validität für die Studienfächer Jura und Wirtschaftswissenschaften mit korrigierten Validitäten von .38 und .56. Die weiteren Studienfächer waren Psychologie (.46), Medizin (.45), Geisteswissenschaften und Lehrämter (.46), Mathematik und Naturwissenschaften (.47).

Prognostische Validität verschiedener Schulfächer:

Zusätzlich wurde die Validität der verbreitetsten Schulfächer errechnet:

	N	k	$\bar{\varrho}$	σ^2	%	$\Delta_{crit}(\bar{\varrho})$
Gesamtprädiktor/ -kriterium	26867	75	.456	.005	64	.317–.595
Einzelprädiktoren: Einzelne Schulfächer						
Deutsch	4046	17	.270	.002	83	.182–.358
Englisch	2400	10	.212	.009	46	.026–.398
Französisch	2341	7	.278	.013	58	.054–.501
Latein	2609	10	.226	.005	58	.087–.365
Mathematik	4242	18	.344	.011	47	.138–.549
Physik	4030	16	.307	.007	57	.143–.471
Chemie	3513	14	.266	.009	48	.080–.452
Biologie	3444	13	.193	.009	45	.007–.379
Erdkunde	2581	10	.238	.003	70	.131–.345
Geschichte	2719	11	.271	-.0002	101	.271–.271
Religion	1748	6	.217	.009	44	.031–.403
Musik	1855	7	.173	.008	46	-.002–.348
Kunst	2451	9	.143	.003	70	.036–.250
Sport	1542	5	.069	-.0005	109	.069–.069
Einzelkriterien: Vorexamen vs. Hauptexamen						
Vorexamen	9950	20	.446	.010	43	.250–.642
Hauptexamen	5266	30	.434	.012	50	.219–.649

N Größe der Gesamtstichprobe der jeweiligen Analyse
k Anzahl der unabhängigen Stichproben in der jeweiligen Analyse
$\bar{\varrho}$ Mittlere korrigierte Validität
σ^2 Korrigierte Varianz
% Prozentualer Anteil der Gesamtvarianz, der durch alle Artefakte erklärt wird
Δ_{crit} Konfidenzintervall mit p = 95 % um $\bar{\varrho}$

Abbildung 2: Validitäten für verschiedene Schulfächer in der Studie von Baron-Boldt, Schuler und Funke (1988, S. 82).

Es zeigte sich, dass unter den Einzelfächern der Mathematiknote (.34) als Prädiktor für Studienerfolg die höchste Prognosequalität zukommt, gefolgt von Physik (.31), Französisch (.28), Deutsch (.27), Chemie (.27), Geschichte (.27), Erdkunde (.24), Latein (.23), Religion (.22), Englisch (.21), Biologie (.19), Musik (.17), Kunst (.14). Die geringste Validität erreichte die Sportnote (.07).

Die Aufstellung zeigt, dass keine der Einzelnoten allein die Validität der Abiturgesamtnote erreicht. Zu bedenken ist jedoch, dass hier nicht die prognostische Validität der Schulfächer für einzelne Studiengänge berechnet wurde. So ist durchaus zu erwarten, dass beispielsweise die Sportnote eine durchaus gute prognostische Validität für Sportstudenten besitzen kann.

Erfahrungen aus der Personalpsychologie:

Zur Berufseignungsdiagnostik wird eine Vielzahl eignungsdiagnostischer Verfahren eingesetzt. Diese lassen sich nach verschiedenen Gesichtspunkten klassifizieren.

Schuler (2001) unterscheidet nach 3 grundsätzlichen Methodischen Ansätzen, dem Eigenschafts- bzw. Konstruktansatz, dem Simulationsansatz und dem biographische Ansatz. Mit dem Eigenschaftsansatz werden relativ stabile (homogene) Merkmale erfasst (z.B. Gewissenhaftigkeit oder sprachgebundene Intelligenz), diese werden meist mit psychologischen Tests erfasst. Der Simulationsansatz orientiert sich hingegen an den (heterogenen) Anforderungen der beruflichen Aufgabe, es soll also Verhalten erfasst werden, dass in ähnlicher Form am Arbeitsplatz gezeigt werden soll. Die typische Erhebungsform ist hier die Arbeitsprobe. Bei diesem Ansatz bleiben die hinter dem Verhalten stehenden Eigenschaften (Persönlichkeitsmerkmale) weitestgehend unbestimmt.

Als dritter Ansatz steht der biographische Ansatz. Er bedient sich meist biographischer Fragen, entweder im Interview oder im Fragebogen.

Im Rahmen dieser Arbeit sind besonders die konstruktorientierten Verfahrung von Interesse, da sie diejenigen sind, die auch bei der Studierendenauswahl zum Einsatz kommen sollen. Daher werden im folgenden Aspekte dieser Verfahrensklasse dargestellt.

Nach Schuler (1998) werden in der Personalauswahl vor allem folgende Gruppen von Tests eingesetzt:

- Allgemeine Intelligenztests
- Tests spezifischer kognitiver Fähigkeiten
- Tests der Aufmerksamkeit und Konzentration
- Tests sensorischer und motorischer Leistung
- sonstige Leistungstests (etwa Wissens- oder Rechtschreibungsprüfungen)
- Persönlichkeitstests

Von diesen in der Personalauswahl eingesetzten Gruppen von Test, kämen für die Studierendenauswahl die Allgemeinen Intelligenztests, die Tests spezifischer kognitiver Fähigkeiten, Tests der Aufmerksamkeit und Konzentration sowie die sonstigen Leistungstests in Frage. Die Tests sensorischer und motorischer Leistung hingegen dürften nur in Ausnahmefällen für die Auswahl von Studierenden nützlich sein, etwa für die Auswahl von Sportstudenten. Die Klasse der Persönlichkeitstests kann, wie in der Einleitung dargelegt, derzeit für die Hochschulzulassung nicht verwendet werden.

Tests aus diesen Klassen, die nach Brambring (1983) am häufigsten in Unternehmen, Behörden und ähnlichen Einrichtungen Verwendung finden sind:

- der Intelligenz-Struktur-Test (Amthauer)
- das Leistungs-Prüf-System (Horn)
- der Aufmerksamkeits-Belastungstest d2 (Brickenkamp)
- der Mechanisch-Technischer Verständnistest (Lienert)
- der Differentieller Wissens-Test (Jäger)
- das Freiburger Persönlichkeitsinventar (Fahrenberg, Selg & Hampel)

Zur Validität dieser Eignungsdiagnostischen Verfahren gibt es mehrere Metaanalysen, die Schmidt und Hunter (1998) zusammenfassend wiedergeben. Das Kriterium in den Metaanalyse war eine Leistungsbeurteilung der Arbeit. Die höchsten Validitäten erreichten Arbeitsproben (.54), gefolgt von allgemeinen kognitiven Fähigkeitstests (.51), strukturierten Einstellungsgesprächen (.51), Fachkenntnistests (.48), Probezeit (.44), Integrity Tests (.41), unstrukturierten Einstellungsgesprächen (.38), Assessment Centern (.37), biographischen Daten (.35) und Gewissenhaftigkeitstests (.31). Am schlechtesten prognostizierten die Interessen (.10) der Arbeitnehmer ihre Leistung.

Die hohen Validitäten werden jedoch teilweise dadurch erreicht, dass die einzelnen Verfahren ähnliches erfassen. Dementsprechend fällt die inkrementelle Validität der einzelnen Testverfahren geringer aus. Als ersten Prädiktor wählten Schmidt und Hunter die allgemeine kognitiven Fähigkeitstests. Als zweiten Prädiktor jeweils ein anderes Testverfahren. Den höchsten Validitätszuwachs erzielten die Integrity – Tests, die für die Hochschulzulassung, zumindest in Deutschland, derzeit belanglos sind. Ein strukturiertes Einstellungsgespräch verzeichnete immerhin noch eine inkrementelle Validität von .12, ebenso die Arbeitsproben. Interessant im Rahmen der Hochschulzulassung sind insbesondere noch die Fachkenntnistests, mit einer inkrementellen Validität von .07.

Anzumerken zu der zusammenfassenden Darstellung von Schmidt und Hunter ist jedoch, dass diese nicht ganz unproblematisch ist. So haben die Autoren bei konkurrierenden metaanalytischen Befunden stets die höchste gefundene in ihre Übersicht übernommen. Zudem konnten einige Randvariablem nicht parallelisiert werden, da die Koeffizienten aus unterschiedlichen Metaanalysen stammten. (Schuler, 2001)

Neben diesen Tests entwickeln viele Organisationen eigene Testverfahren, wie z.B. den Berufswahltest (siehe z.B. Engelbrecht, 1994) der Bundesanstalt für Arbeit.

Neben standardisierten Tests zur Erfassung der allgemeinen kognitiven Fähigkeit werden oft auch spezielle Verfahren für bestimmte Berufsgruppen eingesetzt. Diese sollen genau jene Faktoren erfassen, die für eine spezielle berufliche Tätigkeit von Bedeutung sind. Insbesondere bei Berufen die weniger komplex sind könnte hier ein spezielles Verfahren von Vorteil sein. Für komplexe Berufe, mit heterogenen Arbeitsanforderungen, liefern hingegen allgemeine Fähigkeiten einen besseren Prädiktor, schon weil es schwer wäre, für

diese Berufe ein spezielles Verfahren zu entwickeln. Dieses spiegelt sich in einer Metaanalyse von Hunter und Hunter (1984) wieder, in welcher die allgemeine Kognitive Fähigkeit für Führungspositionen eine kriteriumsbezogene Validität von .58, für Arbeitsplätze mit geringer Vorbildung von .40 und für Arbeitsplätze ohne Vorbildung von nur noch .23 erreichte.

Testverfahren:

Vorteile von Testverfahren in der Studierendenauswahl sind ihre vergleichsweise hohe Objektivität und Reliabilität, aber auch ein im Vergleich zu Auswahlgesprächen geringerer Aufwand. Jedoch ist auch für Testverfahren der Aufwand für die Erstellung und ständige Anpassung der Tests sehr aufwendig und lohnt sich nur bei einer großen Bewerberzahl.

Studierfähigkeitstests (Aptitude Tests):

Es kann zwischen allgemeinen Studierfähigkeitstests und studienfachspezifischen Fähigkeitstests unterschieden werden. Allgemeine Studierfähigkeitstests sollen kognitive Fähigkeiten messen, die für alle Studiengänge wichtig sind. Studienfachspezifische Tests hingegen testen Fähigkeiten, die für die Bewältigung bestimmter Studiengänge oder Studienfelder bedeutsam sind. Beide Testverfahren besitzen eine hohe Objektivität und gelten als kaum trainierbar und sind hoch objektiv. Verwendung finden diese Tests vor allem in den USA, in Großbritannien und in Schweden. Beide Formen sind in Deutschland nicht sehr verbreitet, bisher überwiegend nur an einigen privaten deutschen Hochschulen. Die Ergebnisse der Studierfähigkeitstests stellen einen guten Prädiktor für den Studienerfolg dar. Die Prognosegenauigkeit lässt sich jedoch durch Hinzunahme von Schulnoten und Kenntnistest weiter erhöhen. (Deidesheimer Kreis, 1997).

Studienfachspezifische Tests:

Studienfachspezifische Tests sind so konzipiert, dass sie die spezifischen Anforderungen einzelner Studienfächer berücksichtigen. Das sicherlich prominenteste und verbreitetste Verfahren ist der Test für die medizinischen Studiengänge (TMS), der jedoch eingestellt wurde. Er diente der Studierendenauswahl in den Studiengängen Medizin, Tiermedizin und Zahnmedizin. Das Verfahren enthält neun Untertests (Muster zuordnen, Medizinisch-naturwissenschaftliches Grundverständnis, Schlauchfiguren, Quantitative und formale Probleme, Konzentriertes und sorgfältiges Arbeiten, Figuren lernen, Fakten lernen, Textverständnis und Diagramme und Tabellen). Der TMS sollte möglichst Merkmale testen, die nicht durch die Abiturnote erfasst werden. Dennoch zeigten Validierungsstudien Zusammenhänge zwischen beiden Maßen, im Studiengang Medizin eine Korrelation von r = 0,44 zwischen dem TMS-Ergebnis und Leistungen in der Ärztlichen Vorprüfung. Diese lagen etwas höher als zwischen der Abiturnote und der Leistungen in der ärztlichen Vorprüfung (r = 0,38). Eine Kombination von Abiturnote und TMS-Wert ergab einen

Zusammenhang von r = 0,51, es gab also einen inkrementellen Zuwachs. In der der Tier- und Zahnmedizin jedoch lagen die Korrelationen des TMS unter denen der Abiturnote. (vgl. Köller und Baumert, 2002)

Allgemeine Studierfähigkeitstests

Allgemeine Studierfähigkeitstests sollen Fähigkeiten erfassen, die zentrale Voraussetzungen für beinahe alle universitären Fächer sind. Sie bestehen meistens aus einem verbalen und einem quantitativen Teil.

Beispiele für diese Testform sind der in den USA weit verbreitete Scholastik Aptitude Test (SAT), oder der in Deutschland vormals verwendete Auswahltest der Studienstiftung (ATS). Ein dem ATS ähnlicher Test, ist der Test der akademischen Befähigung (TAB). Der SAT ist ein Multiple-Choice Test und wird von den meisten Amerikanischen Colleges und Universitäten gefordert. Die Testergebnisse dienen oft einer Vorauswahl der Bewerber. (vgl. Deidesheimer Kreis, 1997).

Hezlett et al. (2001, zitiert nach Robins et al. 2004) führte eine Metaanalyse zur prognostischen Validität des SAT zur Vorhersage des Studienerfolgs durch. Als Prädiktor dienten die SAT Testergebnisse, als Kriterium die Durchschnittsnoten (GPA) der Studierenden im College. Nach einer Korrektur der Messbereichseinschränkungen und Korrekturen für die Ungenauigkeit der Messung des Kriteriums erzielten sie Validitäten von .40 und .50. Die Korrekturen der Messbereichseinschränkungen wurden durchgeführt, da die Varianz der SAT Testergebnisse in den Stichproben der Studierenden kleiner ist, als in der Bewerberpopulation. Der korrelative Zusammenhang zwischen SAT Testergebnis und Studienerfolg unterschätzt folglich die prognostische Validität des SATs, dieses soll durch die Korrekturen ausgeglichen werden.

Ähnliche Ergebnisse fanden Robins et al (2004) in einer Metaanalyse von 109 amerikanischen Einzelstudien. In dieser Analyse erzielten die ACT/SAT Testergebnisse als Prädiktoren für Studienerfolg eine Validität von .37. Dieses Ergebnis ist zwar etwas niedriger als das von Hezlett et al., jedoch ist zu bedenken, dass Robins et al keine Korrekturen für Messbereichseinschränkungen vornahmen. Dies könnte das etwas niedrigere Ergebnis erklären.

Kenntnistests (Achievement Tests):

Fachkenntnisse gehören zu den wichtigsten Prädiktoren von beruflichem Erfolg. Kenntnistest sind weit verbreitet und werden beispielsweise in den USA, Japan, China, Griechenland, Israel, Südkorea und in der Türkei bei der Hochschulzulassung verwendet (Deidesheimer Kreis, 1997). Sie sollen den Wissensstand im Allgemeinen oder in bestimmten Bereichen, die in einem Zusammenhang mit dem Studienfach stehen, erfassen. Es kann zwischen schulfach- und studienfachspezifischen Kenntnistests unterschieden

werden. Bei der Testung werden überwiegend Multiple-Choice-Verfahren verwendet, in einigen Tests sind jedoch auch Fragen mit offenen Antworten zu bearbeiten.

Validierungsstudien zu schulfach- und studienfachspezifischen Kenntnistests haben gezeigt, dass diese generell zur Feststellung der Studierfähigkeit geeignet sind, in ihrer Prognosekraft aber hinter den gemittelten Abschlussnoten (z. B. Gesamtnote im Abitur) zurückbleiben (Deidesheimer Kreis, 1997).

Schulfachbezogene Kenntnistests

Schulfachbezogene Kenntnistests dienen der Überprüfung schulischen Wissens. Sie können Mängel des Prädiktors Schulnoten ausgleichen, indem sie eine bessere Vergleichbarkeit von Kenntnissen ermöglichen. In den Ländern in denen solche Tests eingesetzt werden, liegen überwiegend Tests für die Bereiche Muttersprache, Mathematik, Naturwissenschaften (Biologie, Chemie, Physik), Sozialwissenschaften (Politik, Geographie, Geschichte, Ethik, Ökonomie) und Fremdsprachen vor (Deidesheimer Kreis, 1997). Ihnen liegt die Annahme zugrunde, dass die erfolgreich erworbenen schulischen Kenntnisse maßgeblich für den Studienerfolg sind.

Als Beispiel eines solchen Tests kann der in Japan verwendete National Center Test (NCT) aufgeführt werden. Dieser Test dient der Überprüfung des schulischen Leistungsstandes bzw. des individuellen Wissensstandes, den ein Schüler in der Oberstufe erreicht hat. Dieser Test ist ein reiner Multiple-Choice-Test und soll eine einheitliche und objektive Schätzung der Kenntnisse der Bewerber ermöglichen. Zusätzlich zu diesem Test verwenden die Hochschulen jedoch zusätzliche Auswahlverfahren, die auf das jeweilige Hochschulprofil zugeschnitten sind. (Deidesheimer Kreis, 1997)

Studienfachspezifische Kenntnistests

Kenntnistests, die hingegen studienfachspezifische Kenntnisse prüfen, sind nicht dem Lehrinhalten bestimmte Schulfächer zugeordnet. Sie sollen vielmehr die Voraussetzungen testen, die für die Aufnahme eines bestimmten Hochschulstudiums für erforderlich gehalten werden. In den USA zählt der Medical College Admission Test zu diesem Test-Typus.

Empirische Befunde:

Schriftliche Fachkenntnistests erreichten in einer Metaanalyse von Dye, Reck und McDaniel (1993, zitiert nach Schuler 2001) eine durchschnittliche korrigierte Validität von r=.45 (Kriterium Berufsleistung). Dabei fanden sie zwei Moderatoren der Validität, Aufgabenkomplexität und Ähnlichkeit zwischen Test und Arbeitstätigkeit. Bei geringerer Komplexität war die Validität niedriger (.39 vs. .46). Noch stärker wirkte sich der Moderator Ähnlichkeit aus. War die Ähnlichkeit zwischen Test und Aufgabe gering, lag die Validität nur bei .35, bei hoher Ähnlichkeit hingegen bei .62. Folglich sollten Fachleistungstests der Arbeitstätigkeit möglichst ähnlich sein.

Fremdsprachentests:

Fremdsprachentests sind eine besondere Form der Kenntnistests. Sie werden beispielsweise in Japan zur Studierendenauswahl verwendet. Die Tests beinhalten üblicherweise Aufgaben zum Wortschatz, zur Grammatik, zum Hör- und zum Leseverständnis. Amerikanische Universitäten verlangen von nichtamerikanischen Studenten das erfolgreiche Abschneiden im Test of English as a Foreign Language (TOEFL). Um an einer amerikanischen Universität zum Studium zugelassen zu werden, muss man in diesem Test bestimmte Schwellenwerte erreichen. Das Erreichen von Cutt-Off-Werten in diesem Test ist ein notwendiges, aber nicht hinreichendes Kriterium für die Aufnahme an einer Universität. Der TOEFL besteht aus drei Untertests (Hörverständnis, Wortschatz und Leseverständnis, Rechtschreibung und Grammatik). Validierungsstudien des Educational Testing Service in Princeton, New Jersey, wo der TOEFL entwickelt wurde und ständig modifiziert wird, haben moderate Zusammenhänge (um r = 0,30) mit Studienleistungsindikatoren ergeben (Köller & Baumert, 2002). Fremdsprachentests dienen in erster Linie der Auswahl ausländischer Studierende. Sie stellen eine spezielle Form der Kenntnistests dar. Denkbar wäre jedoch auch alle Studienbewerber eines solchen Verfahrens zu unterziehen, da insbesondere Englischkenntnisse im universitären Bereich zunehmend an Bedeutung gewinnen.

Fazit:

Es wurden verschiedene Prädiktoren des Studienerfolgs dargestellt. Es zeigte sich, dass die Abiturnote eine zufriedenstellende prognostische Validität aufweist. Kein anderes Testverfahren konnte durchweg gleich hohe Ergebnisse erreichen. Aus diesem Grund sollte sie weiterhin in Auswahlverfahren einbezogen werden. In den dargestellten Studien zeigte sich, dass keine der Einzelnoten allein die Validität der Abiturgesamtnote erreicht. Bezüglich der Einzelnoten liegt ein weiterer Forschungsbedarf vor. Hilfreich wären Studien, die systematisch einzelne Schulfächer sowie unterschiedliche Gewichtungen eben dieser Fächer als Prädiktoren des Studienerfolgs untersuchen.

Schulfachbezogene Kenntnistests erheben überwiegend redundante Informationen, ihr Einsatz in Auswahlverfahren ist deshalb fragwürdig. Es bleibt aber letztendlich eine politische Abwägung, ob der hohe Aufwand den entstehenden Nutzen rechtfertigen kann. Die Metaanalyse von Dye et al. (1993) zeigte, dass die Ähnlichkeit zwischen den im Test erfassten Kenntnissen und den im Beruf geforderten, die Beziehung zwischen Prädiktor und Kriterium moderierte. Aus diesem Grund sind auch bei Hochschulzulassungen studienfachspezifische Kenntnistests vorzuziehen, die sich stark an den Anforderungen des jeweiligen Studiums orientieren.

Eine sinnvolle Ergänzung der Abiturnoten stellen auch Studierfähigkeitstests dar. Es ist abzuwägen, ob der höhere Aufwand studienfachspezifischer Tests, im Vergleich zu allgemeinen Studierfähigkeitstests gerechtfertigt ist.

Bei der Auswahl der Zulassungsinstrumente sollte stets der inkrementelle Zuwachs an erklärter Varianz, den ein Verfahren erreicht, im Auge behalten werden. Hierzu sind weitere Forschungsarbeiten wünschenswert. Studien aus dem Bereich der Personalauswahl zeigten diesbezüglich, dass viele Auswahlverfahren ähnliches erfassen. Deshalb sollte stets eine Evaluierung der eingesetzten Verfahren stattfinden, da nur so vermieden werden kann, dass zuviel redundante Informationen erfasst wird, die dem Ziel, ein auch verhältnismäßig ökonomisches Auswahlverfahren zu entwickeln, entgegenwirken würde.

Literatur

Althoff, K. (1986). Zur Aussagekraft von Schulzeugnissen im Rahmen der Eignungsdiagnostik. *Psychologie und Praxis*, 30(2), 77-85

Baron-Boldt, J., Schuler, H. & Funke, U.(1988). Prädiktive Validität von Schulabschlußnoten: Eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*, 2, 79-90.

Brambring, M. (1983). Spezielle Eignungsdiagnostik. In K. J. Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie*: Themenbereich B, Methodologie und Methoden: Serie 2 Psychologische Diagnostik, Band 2 Intelligenz- und Leistungsdiagnostik. Göttingen: Hogrefe.

Deidesheimer Kreis (1997). *Hochschulzulassung und Studieneignungstests: studienfeldbezogene Verfahren zur Feststellung der Eignung für Numerus-Clausus- und andere Studiengänge.* Göttingen, Zürich: Vandenhoeck und Ruprecht.

Dye, D. A., Reck, M., & McDaniel, M. (1993). The validity of job knowledge measures. *International Journal of Selection and Assessment*, 1, 153-162

Engelbrecht, W. (1994). Computerunterstützte berufsbezogene Testauswertung im Dienst der Berufsberatung. *Zeitschrift für Arbeits- und Organisationspsychologie*, 38.

Hezlett, S. A., Kuncel, N. R., Vey, M. A., Ahart, A. M., Ones, D. S., Campbell, J. P., et al. (2001,April).*The predictive validity of the SAT: A meta-analysis.* Paper presented at the16th annual conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternate predictors of job performance. *Psychological Bulletin*, 96, 72-98

Köller, O. & Baumert, J. (2002). Das Abitur - immer noch eingültiger Indikator für die Studierfähigkeit? *Aus Politik und Zeitgeschichte*, B 26

Rindermann, H. & Oubaid, V. (1999). Auswahl von Studienanfängern durch Universitäten – Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 20(3), 172-191.

Robbins, S. B., et al. (2004). Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychological Bulletin*, 130(2), 261-288 .

Schmidt, F.L. & Hunter, J. E. (1998). Messbare Personmerkmale: Stabilität, Variabilität und Validität zur Vorhersage zukünftiger Berufsleistung und berufsbezogenen Lernens. In M. Kleinmann & B. Strauss (Hrsg.), *Potentialfeststellung und Personalentwicklung* (S. 16-43). Göttingen: Hoegrefe.

Schuler, H. & Höft, S. (2001). Konstruktorientierte Verfahren der Personalauswahl. In: Schuler (2001, 93-133).

Schuler, H. (2001). *Lehrbuch der Personalpsychologie.* Göttingen: Hoegrefe.

Schuler, H. (1998). Noten und Studien- und Berufserfolg. In: Rost, D. H. (Hrsg.) (2001, 501-507). *Handwörterbuch Pädagogische Psychologie*. 2. Aufl. Weinheim: Psychologie Verlags Union.

Schuler, H (1998). *Psychologische Personalauswahl*. Göttingen: Hogrefe.

Trost, G. & Bickel, G. (1979). *Studierfähigkeit und Studienerfolg*. München: Minerva.

Das Auswahlinterview

Heimo Düvel

Einleitung

Die Psychologische Berufseignungsdiagnostik besteht im Bemühen, Zusammenhänge zwischen menschlichen Merkmalen und beruflichen Erfolg zu entdecken und Methoden zu entwickeln, um beides zu messen und zueinander in Beziehung zu setzen. Es lässt sich zeigen, dass die sachgemäße Anwendung dieser Methoden bessere Prognosen des Berufserfolgs ermöglicht als jede andere Vorgehensweise (Schuler, 2000).

Beruflicher Erfolg hängt von vielem ab, z.B. dem familiären und sozialen Hintergrund, welche die Einstellungen und Erwartungen eines Menschen prägen oder die Ausbildung, welche die Grundlage für weitere Entwicklungsmöglichkeiten schafft und vieles mehr. Gleichzeitig kann beruflicher Erfolg inhaltlich unterschiedlich definiert werden, z.B. durch Leistung, durch Sinnerleben, durch psychische und physische Gesundheit, usw. Um den persönlich definierten Erfolg zu erlangen, sollte es letztendlich zu einer „Passung“ zwischen Person und Tätigkeit kommen. Die Auswahl einer Studienrichtung ist somit gesehen, ein wichtiger Schritt auf dem Weg zu diesem Erfolg.

Das Interview als Methode zur Personalauswahl, trägt zusammen mit anderen Eignungsdiagnostische Verfahren dazu bei, eine breitere Informationsbasis zu schaffen um diese „Passung“ zwischen Person und Tätigkeit zu optimieren. Ziel dieses Beitrags ist es wichtige Aspekte des Eignungsdiagnostischem Interview zu erörtern.

Eignung und Eignungsdiagnostik

Die Berufseignung, verstanden als Erfolgswahrscheinlichkeit, heißt im Prinzip immer „Eignung wessen wofür“ und somit sind beides, die Zielposition und auch die Zielgruppe zu bestimmen. Die Anforderungen sind zu ermitteln, die diese Tätigkeiten an ihre Inhaber stellen, und es sind daraus die erforderlichen Eignungsmerkmale der Positionsinhaber abzuleiten. Um die Ausprägungen der Eignungsmerkmale im Einzelfall zu messen und zu vergleichen, stehen eignungsdiagnostische Verfahren zu Verfügung oder werden konstruiert (Schuler & Marcus, 2001).

Die Eignungsdiagnostik hat zum Ziel, Person und Tätigkeit auf den Ebenen Leistungspotential, Befriedigungspotential und Entwicklungspotential zu vergleichen um somit eine Passungsentscheidung abzuleiten. Voraussetzung für die Diagnostik ist also eine möglichst detaillierte Anforderungsanalyse.

Eignungsdiagnostische Instrumente lassen sich nach drei grundsätzlichen methodischen Ansätzen unterscheiden (Schuler & Marcus, 2001). Mit dem **Eigenschaftsansatz** werden

Merkmale erfasst, die als relativ stabil angenommen werden. Typische Messverfahren zur möglichst homogenen Erfassung der Merkmale sind psychologische Tests. Zielsetzung des **Simulationsansatzes** ist die Erfassung solchen Verhaltens, das in ähnlicher Form am Arbeitsplatz gefordert wird mit typischer Erhebungsform die Arbeitsprobe. Der **biographische Ansatz** mit typischer Erfassungsmethode biographische Fragen, zielt auf die Evaluation von Ergebnissen, bzw. vergangenen Verhalten ab, um somit möglichst genau zukünftiges Verhalten vorauszusagen. In schriftlicher Form ist dies der biographische Fragebogen und in mündlicher Form ist dies das Eignungsinterview.

Die Verschiedenheit der diagnostischen Ansätze bedeutet für die praktische Eignungsdiagnostik, dass für komplexe Anforderungssituationen meist ein multiples Verfahren angemessen ist, um somit aus den verschiedenen Ansätzen sequentiell Informationen anzureichern. Bevor auf die speziellen Aspekte des Eignungsinterviews eingegangen wird, soll zunächst kurz die Anforderungsanalyse als Voraussetzung für die Diagnostik beschrieben werden.

Arbeits- und Anforderungsanalyse

Anforderungen setzen sich aus tätigkeitsrelevanten oder aus personenrelevanten Merkmalen zusammen. Westhoff (2004) bietet eine Verhaltensgleichung, die Verhalten als Funktion von Gruppen von Bedingungen und deren Wechselwirkungen beschreibt. Darin setzen sich nichtpsychologische Anforderungen zusammen aus Anforderungen aus der Umgebung sowie aus Anforderungen an den Organismus, während psychologische Anforderungen sich in kognitive, emotionale, motivationale und soziale Anforderungen einteilen lassen. Die Verhaltensgleichung bietet somit eine Struktur für das kohärente darstellen der Anforderungen.

Nur eine Untermenge der Begriffe, die Anforderungen einer bestimmten Tätigkeit bezeichnen, ist wissenschaftlich beschrieben. Wissenschaftliche Konstrukte für die Personalauswahl zeichnen sich also dadurch aus, dass man die zugrunde liegenden Überlegungen empirisch geprüft hat und somit erhöhen diese Konstrukte das genaue Bestimmen von Ausprägungen bestimmter Anforderungen.

Arbeitsanalyse

Die Arbeitsanalyse ist eine Methode der Identifizierung der an einem Arbeitsplatz oder in einem Beruf auszuführenden Aufgaben und Tätigkeiten. Die Ausführungsbedingungen sowie die psychischen, physischen und sozialen Bedingungen des Umfeldes und die relevanten Organisationsmerkmale werden beschrieben (Reimann, 2004). Als Informationsquellen dient alles, was Aufschluß über Art, Bedingungen, Konsequenzen der Tätigkeit als auch über Leistungskriterien geben kann. Methoden der Arbeitsanalyse sind Expertenbefragung, Beobachtung, Interviews, Fragebogen (z.B. Fragebogen zur Arbeitsanalyse von Frieling & Hoyos, 1978) sowie die Ausführung der Tätigkeit durch den Analytiker.

Es wird zwischen folgenden Anforderungen unterschieden (Reimann, 2004):

a) Die **Fachkompetenz** bezieht sich auf die zu erreichenden Ziele bzw. Leistungsergebnisse der Stelle und die jeweils nötigen Ausbildungen, Fachkenntnisse und Berufserfahrungen.

b) Die **Methodenkompetenz** beinhaltet die notwendigen Aktivitäten und das Anwenden fachlicher Kenntnisse um Problemlöseprozesse zu gestalten.

c) Die für die Aufgabenerfüllung erforderlichen **Fähigkeiten im Umgang mit anderen** Menschen wie Kommunikationsfähigkeit, Teamfähigkeit, Konflikt- und Kritikfähigkeit, der Aufbau und die Pflege von Beziehungsgeflechten, usw.

d) Die **Persönlichkeitskompetenz** bezieht sich auf die zur Leistungserfüllung notwendigen Personenmerkmale wie z.B. Engagement, Intelligenz, analytische Fähigkeiten, Lernbereitschaft, Kreativität, Stabilität, usw.

e) Ausgehend von den Zielen und Strategien des Unternehmens spiegelt sich die **Unternehmenskompetenz** in der Identifikation mit der Unternehmenskultur, der Vertrautheit mit den Unternehmensprozessen und der Loyalität gegenüber dem Unternehmen wider.

Anforderungsanalyse

Unter der Anforderungsanalyse wird die Ermittlung von personrelevanten psychischen Voraussetzungen für den zu besetzenden Arbeitsplatz, das Aufgabenfeld, die Ausbildung bzw. das Studium verstanden, für das die Eignung einer zu beurteilenden Person festgestellt werden soll (Reimann, 2004).

Es wird zwischen folgenden Anforderungen unterschieden (Reimann, 2004):

a) Eigenschaftsanforderungen (z.B. Fähigkeiten und Interessen);

b) Verhaltensanforderungen (z.B. Fertigkeiten und Gewohnheiten);

c) Qualifikationsanforderungen (z.B. Kenntnisse und Fertigkeiten);

d) Ergebnisanforderungen (z.B. Problemlösungen und Qualitätsstandards).

Methodische Zugänge zur Bestimmung von Anforderungen

Zur Bestimmung von Anforderungen stehen grundsätzlich drei methodische Zugänge zur Verfügung (Reimann, 2004). Bei der **erfahrungsbasierten-intuitiven Methode** basiert die Einschätzung der Anforderung auf der erfahrungsgeleiteten Beurteilung der Tätigkeit, den erforderlichen Arbeitsmitteln und Arbeitsgegenständen, den Umweltbedingungen, den Qualifikations- und Weiterbildungserfordernissen usw. Diese Methode ist bei ausreichender Erfahrung des Beurteilers sinnvoll und effektiv. Bei der **arbeitsanalytisch-empirischen Methode** kommen teil- oder vollstandardisierte Fragebogen, Checklisten und

Arbeitsanalyseverfahren zum Einsatz. Es werden Tätigkeitselemente ermittelt, die in Personenanforderungen übersetzt werden. Letztlich, bei der **personenbezogen-empirische Methode** werden statistische Zusammenhänge zwischen Personenmerkmalen einerseits und Tätigkeitsfolgen (z.B. Belastung, Sättigung, Leistungshöhe, Berufszufriedenheit) anderseits dazu genutzt, die Anforderungen zu bestimmen. Personenmerkmale, die durch Training oder Übung beeinflusst werden, schränken den Einsatz dieser Methode ein.

Das Anforderungsprofil

Das Ergebnis der Arbeits- und Anforderungsanalyse ist das Anforderungsprofil in Form von Verhaltensbeschreibungen. Es enthält alle benötigten und wünschenswerten Voraussetzungen und Kompetenzen einer Person für den zu besetzenden Arbeitsplatz, Aufgabenfeld, einer Ausbildung oder einem Beruf, einschließlich der Merkmale die für die berufliche Zufriedenheit wichtig sind. Ausgehend von diesem Anforderungsprofil werden nun ausgewählte Verfahren der Eignungsdiagnostik eingesetzt um die Passung zwischen Bewerber und Stelle zu bewerten.

Der Biographische Ansatz

Die Grundannahme des biographischen Ansatzes ist, dass menschliches Verhalten in gewissem Ausmaß über lange Zeiträume stabil bleibt.

Das gemeinsame Merkmal von Bewerbungsunterlagen wie Schul- und Ausbildungszeugnisse, der Lebenslauf, Referenzen, usw. ist, dass sie über die Vergangenheit des Bewerbers informieren. Darin spiegeln sich Ereignisse, Verhalten und Leistungen wider, aus denen eignungsdiagnostische Prognosen zukünftigen Verhalten abgeleitet werden können. Zum Beispiel zeigen Metaanalysen eine erstaunlich hohe Korrelation für die Beziehung zwischen Schulnoten und Ausbildungserfolg (Schuler, 1998).

Jedoch leiden viele Komponenten von Bewerbungsunterlagen unter einem Standardisierungsdefizit, das sich nicht immer nur auf die Unterlagen selbst bezieht, wohl aber meist auf deren Auswertung. Das Problem dieser klinischen Urteilsbildung besteht vor allem in der mangelnden Reliabilität des Urteils. Im Folgenden wird kurz auf eine der beiden Erfassungsformen biographischer Daten, der biographische Fragebogen, eingegangen, bevor eine detaillierte Beschreibung der zweiten Erfassungsform, das Interview, folgt.

Der Biographische Fragebogen

Die schriftliche Form des Biographischen Ansatzes ist der biographische Fragebogen. Für diesen wurde bis vor kurzen das Prinzip angewandt, einzelne Items anhand ihrer empirisch ermittelten Validität zur Prognose eines Außenkriteriums auszuwählen. Eine Folge dieser streng empirizistischen Vorgehensweise war, dass sich die Methode „biographischer Fragebogen“ von dem Grundaxiom des biographischen Ansatzes, dass nämlich der beste

Prädiktor zukünftigen Verhaltens, das Verhalten in der Vergangenheit sei, teilweise löste. So wurden zum Beispiel Items in den Fragebogen aufgenommen die inhaltlich eher einer Selbsteinschätzung entsprachen. Obwohl durch dieses Vorgehen der Zusammenhang des Items mit dem Kriterium befriedigte, blieb oftmals im Dunkeln, was damit eigentlich gemessen wurde (Schuler & Marcus, 2001).

Heute nähert sich die Methode wieder der Grundannahme des biographischen Ansatzes. So beginnt die Konstruktion eines Fragebogens nunmehr nicht mit der Skalenbildung, sondern mit der Formulierung einzelner Fragen und Antwortmöglichkeiten, die die Bausteine der Skala bilden. Mael (1991) unterscheidet zehn mögliche Attribute zur Kennzeichnung biographischer Items, z.B. historisch – hypothetisch, external – internal, objektiv – subjektiv, aus erster Hand – aus zweiter Hand, usw. Mael's Kernaussage ist das einzig das historische Attribut zur definitorischen Abgrenzung biographischer Daten unerlässlich ist. In den meisten Fällen sollten Items jedoch auch external, objektiv und aus erster Hand sein.

Der Biographische Fragebogen ist die am stärksten formalisierte und wohl auch am intensivsten erforschte Methode zur Erhebung biographischer Daten, deren Effektivität in der psychologischen Eignungsdiagnostik gut abgesichert ist. Allerdings lassen sich biographische Inhalte auch im mündlichen Gespräch so erheben, dass dadurch Aussagen von hoher prognostischer Validität ermöglicht werden.

Das Eignungsinterview

Unter einem Interview als Methode zur Personalauswahl ist eine Gesprächssituation zwischen zwei oder mehreren Personen – Repräsentanten der ausgewählten Organisation einerseits und Stellenbewerber anderseits – zu verstehen, die Gelegenheit zum Austausch bewerbungsrelevanter person-, arbeits- und organisationsbezogener Information bietet und damit als Grundlage für Auswahlentscheidungen seitens der Organisation und der Organisationswahl seitens der Bewerber dient (Schuler & Marcus, 2001).

Einstellungsinterviews sind nach der Auswertung der Bewerbungsunterlagen die verbreitetste Methode der Personalauswahl. Das Gespräch als eignungsdiagnostische Methode genießt hohe Wertschätzung nicht nur seitens der Verwender, sondern auch von seitens der Bewerber. Beide sehen das Gespräch als taugliche Methode an, sich ein Bild vom Gegenüber zu machen und relevante Information als Entscheidungsgrundlage zu gewinnen. Beide schätzen an der Gesprächssituation die Möglichkeit des direkten persönlichen Austauschs und halten die dabei zu gewinnende Hinweise für glaubhafter als die aus schriftlich übermittelter Information.

Interviewer nutzen darüber hinaus Auswahlgespräche als Gelegenheit zur Öffentlichkeitsarbeit. Bewerber schätzen die Möglichkeiten der Situationskontrolle höher ein als bei anderen Auswahlverfahren und nutzen die Situation als Gelegenheit, Feedback zu gewinnen,

sowie als Grundlage ihrer Selbstselektion. Interviewte sehen im Eignungsinterview ein Verfahren, das es ihnen erlaubt, sich selbst möglichst gut darzustellen.

Weiterhin liefert das Interview Informationen, die mit anderen Verfahren nicht oder nur mit unvertretbar hohem Aufwand erhoben werden können. Informationen aus anderen Informationsquellen, wie z.B. Schul- und Arbeitszeugnissen, können adäquater durch die Informationen aus dem Eignungsinterview verstanden und beurteilt werden (Westhoff & Strobel, 2004).

Schließlich bietet das Interview als einziges Auswahlverfahren den Rahmen, Vereinbarungen über den weiteren Auswahlprozess und über die Bedingungen der Zusammenarbeit zu treffen. Dementsprechend gehört das Interview nicht nur zu den von den Verwendern am besten bewerteten Verfahren, sondern wird von den Bewerbern sogar vor allen anderen Auswahlmethoden präferiert.

Der Durchführungsmodus reicht von der völlig freien Gesprächsform über teilstrukturierte bis zu vollstrukturierten Varianten mit standardisierten Abläufen und Fragestellungen. Die Fragestellung beziehen sich insbesondere auf eignungsrelevante Erfahrungen und Ausbildungen, auf Aspekte des Lebenslaufs und deren subjektive Verarbeitung, gelegentlich auch auf persönliche Bereiche wie den des familiären Hintergrunds. In erweiterter Form sind auch Arbeitsproben wie kleine Rollenspiele integriert. Das Interview stellt somit ein Auswahlverfahren dar, das den eigenschaftsorientierten, den simulationsorientierten und den biographieorientierten Ansatz unkompliziert integrieren könnte.

Zur Strukturierung eines Interviews

In der Strukturierung von Auswahlgesprächen, beginnend mit einer fundierten Anforderungsanalyse bis hin zu feststehenden Auswertungsvorschriften, liegt die zentrale Möglichkeit zur Qualitätssicherung von Interviews.

Die Strukturierung von Interviews kann aus einer Vielzahl von Einzelmaßnahmen bestehen. Campion, Palmer und Campion (1997) beschreiben mehrere Strukturierungsmaßnahmen sowie deren Zusammenhang zur Reliabilität, Validität und der Reaktion von Verwendern und Bewerbern in Beziehung stehen. Diese Maßnahmen beziehen sich auf die Inhalts- oder auf die Auswertungsebene.

Maßnahmen der Interviewstrukturierung und ihre Auswirkung auf Reliabilität, Validität und Verwenderreaktionen (Campion et al., 1997; in Schuler & Marcus, 2001)(„+“ bedeutet positiver Effekt, „-“ bedeutet negativer Effekt.

	Reliabilität						**Validität**			**Verwenderreaktion**		
Inhalt	Test-retest	Inter-rater	Kandid. Konsist	Interaktion zw. Kandi-daten	Interne Konsi-stenz	Beurteiler-übersein-stimmung	Anforde-rungs-bezug	Reduz. Defienz	Reduz. Konta-mination	Reduz. Benach-teilg. best. Gruppen	Akzeptanz seitens Kandidaten	Akzeptanz seitens Interviewer
1. Anforderungsanalyse							+	+	+	+	+	+
2. Gleiche Fragen	+	+	+	+				+	+	+		
3. Verzicht auf Hilfen und Nachfragen	+	+	+	+				-	+	+	-	-
4. Bessere Fragen			+		+		+		+	+		+
Längere Interviews	+	+			+			+			-	-
Verzicht auf ergänzende Information	+	+						-	+	+	-	-
Keine Fragen der Kandidaten	+	+	+	+				+	+		-	-
Auswertung												
8. Beurteilung der einzelnen Antowrten oder Verwendung von Einstufungsdimension	+	+			+			+	+			
9. Verankerte Einstufungsskalen	+	+				+	+	+	+	+		+
10. Detaillierte Aufzeichnung	+	+				+	+	+	+	+		-
11. Mehrere Interviewer	+	+		+	+	+		+	+	+	-	
12. Gleiche Interviewer	+			+					+	-		
13. Kein Meinungsaustausch zw. Interviewern	+	-				-			+	+		-
14. Training	+	+	+	+	-	+	+	+	+	+	+	+
15. Statistische Prognose	+	+			+			+	+	+		

Die Strukturierung zieht sich durch die Phasen Planung, Durchführung und Auswertung (Westhoff & Strobel, 2004). Bei mehrfacher Durchführung wird dadurch ein gleicher Ablauf des Interviewprozesses sichergestellt.

A) <u>Planung</u>

Voraussetzung ist ein Anforderungsprofil, in dem differenziert die zur Bewältigung der zukünftigen Arbeitsaufgaben notwendigen Anforderungen mit den dazugehörigen Verhaltensweisen aufgeführt sind. Ein Interviewleitfaden wird erarbeitet. Er enthält alle möglichen Fragen, Einleitungen, Überleitungen, Zusammenfassungen sowie alle zusätzlichen Bereiche.

B) Durchführung

Bei jeder Durchführung müssen alle vorgesehenen Fragen jedesmal gestellt werden. Wichtig ist auch die Reihenfolge und mögliches Nachfragen. Die einzelnen Abschnitte in einem Eignungsinterview können wie im Leitfaden vorgesehen mit den gleichen Überleitungen, Begründungen und Erklärungen eingeleitet werden.

Antworten können vollständig wörtlich oder alle relevanten Informationen können in Stichworten notiert werden. Danach können die Informationen zu einem bestimmten Abschnitt oder einer bestimmten Anforderung nach einer vorher festgelegten Regel beurteilt werden.

C) Auswertung

Die Beurteilung zu den einzelnen Anforderungen werden am besten nach einer expliziten Regel zu einer Gesamtbeurteilung der Informationen aus dem Eignungsinterview zusammengefasst.

Aus Sammelreferaten zeigt sich das die Vorhersagekraft eines Interviews generell mit erhöhter Strukturierung wächst, es jedoch ein Optimum an Strukturierung gibt, über das hinaus keine Validitätssteigerung mehr erfolgt. Bei vollständiger Standardisierung geht die Flexibilität der Interviewführung verloren und kein Eingehen auf den Bewerber ist möglich (Schuler & Marcus, 2001).

Typen strukturierter Interviews

Nachdem sich strukturierte Interviews der freien Gesprächsform hinsichtlich der Validität als grundsätzlich überlegen erwiesen haben, entstanden verschieden konkurrierende Verfahrensformen welche z.B. in Westhoff & Strobel (2004) zusammengefasst sind.

Das Behavior Description Interview (BDI)

Das BDI (Janz, Hellervik & Gilmore, 1986) folgt der Grundannahme, dass Verhalten in bestimmten Situationen in der Vergangenheit der beste Prädiktor für Verhalten in entsprechenden zukünftigen Situationen ist. Daher wird im BDI nach dem Verhalten des Interviewten in tatsächlich erlebten Situationen gefragt. Der Interviewte soll die darin erlebten kritischen Ereignisse schildern. Ein kritisches Ereignis ist ein im Arbeitsalltag typisches und wichtiges Ereignis, in dessen Bewältigung sich gute und weniger gute Stelleninhaber unterscheiden.

Der Vorzug dieses Ausgangsmaterials ist, dass es sich um reale Ereignisse handelt, nicht um Meinungen oder andere unbelegte Äußerungen. Die Sammlung dieser Ereignisse in der gegebenen Organisation erhöht die Anpassung des Verfahrens an die dort gegebenen Bedingungen.

Die kritischen Ereignisse werden zu Leistungsdimensionen gruppiert, die sich im Wesentlichen durch Aufgabenbereiche ergeben. Im Interview werden mehrere Fragen zu jeder dieser Leistungsdimension gestellt und die Antworten protokolliert.

Verschiedene Validierungsstudien ermittelten zufrieden stellende Werte zwischen 0.48 und 0.54 für die Güte des BDI (Westhoff & Strobel, 2004).

Das Situative Interview (SI)

Das SI (Latham, Saari, Pursell & Campion, 1980) geht von der Grundannahme aus, dass sich Menschen entsprechend ihren Zielen und Absichten verhalten. Im SI werden bestimmte zukünftige Situationen in den Fragen vorgestellt, in denen kritische Ereignisse ablaufen, und es wird erfasst wie der Interviewte sich vorstellt, dieses kritische Ereignis zu bewältigen. Das SI geht also wie das BDI von kritischen Ereignissen aus, nur muss der Interviewte diese noch nicht selbst erlebt haben.

Im SI gibt es zur Bewertung der Antworten zu jeder Frage eine fünfstufige Einstufungsskala, bei der die Extreme und der mittlere Wert durch konkretes Verhalten beschrieben sind (verhaltensverankerte Einstufungsskala). Den Kandidaten werden die Antwortverankerungen nicht gezeigt, ebenso wird ihnen nicht der Zusammenhang zu der relevanten Anforderungsdimension mitgeteilt. Validierungsstudien ermittelten Werte zwischen 0.30 und 0.46 für die Güte des SI (Westhoff & Strobel, 2004).

Das Multimodale Interview (MMI)

Das MMI (Schuler, 1992) kombiniert das Vorgehen nach dem BDI und dem SI mit weiteren gut gesicherten Ergebnissen der Forschungen zum Eignungsinterview.

Das Verfahren besteht aus Acht Stufen. Fünf der Stufen dienen der diagnostischen Urteilsbildung, die verbleibenden Drei Stufen erfüllen die Funktion, dem Interaktionsprozeß einen natürlichen Gesprächslauf zu geben.

1. **Gesprächsbeginn**: Kurze informelle Unterhaltung, Bemühen um angenehme und offene Atmosphäre, Skizzieren des Verfahrensablaufs, Keine Beurteilung;
2. **Selbstvorstellung des Bewerbers**: Der Bewerber spricht einige Minuten über seinen persönlichen und beruflichen Hintergrund, seine derzeitige Situation und seine Erwartungen für die Zukunft. Sein Verhalten wird im Hinblick auf anforderungsbezogene Urteilsdimensionen eingestuft.
3. **Berufsorientierung und Organisationswahl**: Es werden standardisierte Fragen zu Berufswahl, Berufsinteressen, Organisationswahl und Bewerbung, bei berufserfahrenen Bewerbern auch zum Fachwissen gestellt. Die Antwortbewertung erfolgt auf verhaltensverankerten Skalen.

4. **Freier Gesprächsteil**: Der Interviewer stellt offene Fragen in Anknüpfung an Selbstvorstellungen und Bewerbungsunterlagen. Die Bewertung erfolgt summarisch.
5. **Biographiebezogene Fragen**: Biographische Fragen werden aus Anforderungsanalysen dimensionsbezogen abgeleitet oder als validierte Fragen aus biographischen Fragebogen übernommen. Die Antwortbewertung erfolgt auf verhaltensverankerten Skalen.
6. **Realistische Tätigkeitsinformation**: Der Interviewer gibt dem Bewerber ausgewogene, bedarfsgerechte Information über Tätigkeit, Arbeitsplatz und Unternehmen. Überleitung zu situativen Fragen. Keine Beurteilung.
7. **Situative Fragen**: Entsprechend dem SI werden dem Bewerber Fragen gestellt zu seinem Verhalten in möglichen zukünftigen erfolgskritischen Situationen. Wie im SI werden die Antworten anhand verhaltensverankerter Einstufungsskalen beurteilt.
8. **Gesprächsabschluss**: Fragen des Bewerbers werden beantwortet. Das weitere Vorgehen wird besprochen. Gegebenenfalls werden bereits Vereinbarungen getroffen.

In bisherigen Studien zur Qualität des MMI streuten die ermittelten Koeffizienten für die prognostische Validität von 0.15 bis 0.57, wobei die meisten Aufgaben zwischen 0,30 und 0,50 lagen. Objektivitätsstudien ermittelten teilweise sehr unterschiedliche Werte für Beurteilungsübereinstimmung in den Gesprächen zwischen 0,32 und 0.90, für den Gesamtwert liegen die Kennwerte bei 0.70 und höher (Westhoff & Strobel, 2004).

Das Entscheidungsorientierte Gespräch

Das EOG ist eine Technologie innerhalb der Entscheidungsorientierten Diagnostik (Westhoff & Kluck, 2003) und eignet sich für alle Arten der diagnostischen Informationsgewinnung mittels Interview.

Das Ziel der EOG-Technologie im Bereich der Eignungsdiagnostik ist, alle gesicherten Vorgehensweisen beim Planen, Durchführen und Auswertung von Eignungsinterviews zusammenzutragen und gleichsam wie einen Werkzeugkasten zur Verfügung zu stellen.

Zur Auswahl von Konzepten und Konstrukten, die man für die Verhaltensvorhersage verwenden kann, dient im EOG die Verhaltensgleichung zur Auswahl von Anforderungen.

Die EOG bietet empirisch geprüfte Regelsysteme:

1. Zu Voraussetzungen für ein erfolgreiches EOG;
2. Zur Planung eines eignungsdiagnostischen Interviews;
3. Zum Grobaufbau eines Leitfadens
4. Zum Feinaufbau eines Leitfadens
5. Zur Formulierung günstiger Fragen
6. Zur Auswertung und Darstellung von Informationen aus Eignungsinterviews

Nach diesen Regeln lässt sich ein vollstrukturiertes Eignungsinterview planen, durchführen und auswerten, das die Grundlagen bietet für eine optimale Gewinnung und Verarbeitung eignungsdiagnostischer Informationen.

Bisherige Studien zur Güte wurden zur Überprüfung der Objektivität durchgeführt und erbrachten Kennwerte bis zu 0,94 für die Beurteilerübereinstimmung (Westhoff & Strobel, 2004).

Der Interviewprozess

Das Eignungsinterview kann von Mitarbeitern aus verschiedenen Bereichen der Organisation gemeinsam geführt werden, was zu höherer Objektivität und besserer Kooperation zwischen den Bereichen führen kann. In diesem Fall ist vorher festzulegen, wer worüber das Gespräch führt und wer die Informationen protokolliert.

Soll ein Eignungsinterview erfolgreich sein und alle mit ihm erhobenen gültigen Informationen zu möglichst geringen Kosten gewinnen, dann (Westhoff & Strobel, 2004):

1. muss es auf gültige, Verhalten beschreibende Anforderungen gegründet sein;
2. müssen alle vom Interviewer zu treffende Entscheidungen bei der Planung, Durchführung und Auswertung nach gültigen vorher explizit festgelegten Regeln getroffen werden;
3. muss der Interviewer in allen relevanten Interviewerverhaltensweisen individuell trainiert und regelmäßig evaluiert werden.

Ein Leitfaden dient dazu diesen Interviewprozess zu formalisieren.

Der Interviewleitfaden

Ein Leitfaden für das Eignungsinterview sollte folgende Punkte berücksichtigen (Westhoff & Strobel, 2004):

1. die notwendigen Informationen für den Beginn des Eignungsinterviews (Begrüßen, Vorstellen, Funktionen der Beteiligten, Ziele und Dauer des Interviews, Übersicht über die Vorgehensweise, Erklärungen zum Umgang mit den zu erhebenden Informationen);
2. alle (möglicherweise) zu stellenden Fragen;
3. die erklärenden Überleitungen zu den einzelnen Abschnitten;
4. notwendige Erklärungen von nicht vermeidbaren Fachbegriffen;
5. Vorschriften, wie der Interviewer bei jeder anstehenden Entscheidung vorzugehen hat.

(Westhoff & Strobel, 2004): Ein Leitfaden ist dann praktisch nützlich wenn:

1. alle Fragen in einem einfachen, klaren und genauen Deutsch formuliert sind;
2. möglichst wenige Fremdwörter vorkommen;
3. notwendige Fachwörter verständlich erklärt werden;
4. alle notwendigen Erklärungen kurz, zutreffend und verständlich sind;
5. notwendige längere Erklärungen in Gesprächsabschnitten vermittelt werden und nicht an einem Stück vorgetragen werden;
6. alle Fragen und Aufforderungen unter einem Gliederungspunkt in der sachnotwendigen Reihenfolge angeordnet sind;
7. in jeder Frage nach konkretem individuellen Fühlen, Denken oder Handeln gefragt wird;
8. nur „günstige“ (siehe unten) Fragen verwendet werden;
9. jede Frage angemessen offen ist;
10. jede Frage angemessen direkt ist;
11. geschlossene Fragen nur als Filterfragen vorkommen.

(Westhoff & Strobel, 2004): Fragen sind in einem Eignungsinterview dann „günstig“ wenn:

1. sie sich auf konkretes individuelles Verhalten beziehen;
2. in einem eindeutigen Bezugsrahmen stehen;
3. sie nur einen Sachverhalt ansprechen;
4. möglichst kurz und treffend sind;
5. nicht suggestiv sind, d.h. nichts ungerechtfertigt als gegeben voraussetzen;
6. möglichst sachlich, d.h. neutral hinsichtlich der Bewertung des erfragten Verhaltens formuliert sind;
7. sie aus Wörtern und Redewendungen bestehen, die möglichst wenig emotional geladen sind;
8. den Kontext als Gedächtnisstütze verwenden;
9. auch die dem Interviewer peinliche Fragen zutreffend formuliert sind;
10. Fragen zur Motivation mit „Warum...?“, „Wieso...?“,“Weshalb...?“, „Was sind die Gründe...?“ ersetzt sind durch die Fragen: „Was finden Sie an gut?“, „Was finden Sie an...weniger gut?“;
11. jede Frage nach Verhalten in hypothetischen Situationen, ersetzt ist.

Voraussetzungen für den Interviewerfolg

Westhoff & Strobel (2004) fassen folgende Punkte zusammen:

1. Der Interviewer hat alle Erwartungen an den Gesprächspartner bei seiner Vorbereitung zugelassen. (Um sich z.B. über seine Sympathie oder Antipathie und deren Anlässe klar zu werden.)
2. Der Interviewer ist sich im Klaren, was er an seinem Gesprächspartner gut findet. (Damit er ihn nicht global zu positiv einschätzt.)
3. Der Interviewer ist sich im Klaren, was er an seinem Gesprächspartner weniger gut findet. (Damit er ihn nicht global zu negativ einschätzt.)
4. Der Interviewer hat konkrete Pläne, wie er mit dem Gesprächspartner umgehen will.
5. Der Interviewer weiß, wie er angemessen mit den von ihm erwarteten Schwierigkeiten umgehen kann.
6. Die Pläne des Interviewers dienen der Sache, um die es geht.
7. Der Interviewer ist gedanklich und gefühlsmäßig richtig auf das Gespräch eingestellt, d.h. es regt ihn daran in der Vorstellung nichts mehr auf.
8. Der Interviewer hat einen übersichtlichen, vollständig und angemessen ausformulierten Leitfaden.
9. Die gewählten Untersuchungstermine sind für alle Beteiligten möglichst günstig.
10. Der Interviewer und der Interviewte haben ausreichend Zeit für das Gespräch vorgesehen.
11. Der Interviewer hat die Umgebung für das Gespräch günstig gestaltet: kleiner Tisch; bequeme Sitzgelegenheiten; für alle Gesprächsteilnehmer gut ablesbare Uhr; Namensschilder; Erfrischungen; kein Telefon oder andere Personen können das Gespräch stören.

Auswertung eines Interviews

Ist das eigentliche Interview vollendet, so beschreiben Westhoff und Strobel (2004) einige wichtige Aspekte der Auswertung, auf die geachtet werden sollte. Am besten ist die vollständige Aufzeichnung des Eignungsinterviews auf Tonträger oder, damit bei Unklarheiten die konkrete Aussage zur Verfügung steht. Wenn ein Interviewer dies dem Interviewten erklärt und ihm zugleich garantiert, dass er die Aufnahme nicht weitergeben wird, dann ist der Interviewte in der Regel bereit, seine vor der Aufnahme unbedingt erforderliche Zustimmung zu geben.

Möglichst umgehend nach dem Eignungsinterview werden die erhobenen Informationen nach vorher festgelegten Regeln zu Antworten auf die Untersuchungsfragen formuliert.

Westhoff & Strobel (2004) setzen folgende Punkte für eine möglichst sachgerechte Auswertung voraus:

1. Die gesamte objektiv registrierte Information wird Schritt für Schritt danach ausgewertet, ob sie in Beziehung zur Fragestellung steht und zu welcher Untersuchungsfrage sie etwas aussagt.
2. Jede Information wird bei jeder Untersuchungsfrage dargestellt, zu deren Beantwortung sie beiträgt.
3. Jede Information wird zutreffend dargestellt.
4. Bei jeder Information ist deutlich, woher sie stammt.
5. Der Kontext, in dem eine Information ursprünglich stand, wird angemessen berücksichtigt.
6. Das Verhalten wird im adverbialen Modus beschrieben.
7. Das Verhalten wird im Imperfekt geschildert.
8. Es werden nur gebräuchliche Wörter verwendet.
9. Es wird Verbalstil verwendet.
10. Es wird die Aktivform verwendet.
11. Jede Information wird in der indirekten Rede dargestellt.
12. Die indirekte Rede wird in der richtigen sprachlichen Form dargestellt. (Die richtige Konjunktivform oder bei jeder Aussage Indikativ mit Relativierung auf die Datenquelle.)
13. Es werden alle möglichen Leser der Gesprächsauswertung bei der Formulierung berücksichtigt.
14. Es wird möglichst sachlich (wenig wertend) formuliert.

Erfolgreiche Eignungsbeurteilung trennt immer die Sammlung von Beobachtungen / Informationen und ihrer Bewertung. Eine fachgerechte Bewertung erfolgt immer erst nach Abschluss der Informationssammlung. Es wird Information für Information daraufhin geprüft, wie gut das beschriebene Verhalten den Anforderungen der zu besetzenden Stelle entspricht. Dabei wird nach vorher festgelegten Entscheidungsregeln entschieden. Die Beurteilung der Bewerberinformationen anhand konkreter und eindeutiger Maßstäbe beugt der Tendenz vor, die Anforderungen unter ungünstigen Bedingungen wie Mangel an fähigen Bewerbern oder Zeitdruck herabzusetzen. Die Einschätzungen der einzelnen Anforderungskriterien auf den Beurteilungsskalen können anschließend durch Summierung kombiniert werden, wobei die vorher festgelegte Gewichtung der Kriterien eingerechnet wird. Als Ergebnis ergibt sich für jeden Bewerber sowohl ein Profil der individuellen Ausprägungen der einzelnen Anforderungen als auch ein Gesamturteil.

Für die Auswertung eines Gespräches durch mehrere Personen ist eine Beurteilungsdiskussion direkt nach dem Interview vorzunehmen, in der die Personen ihre Wahrnehmungen und Meinungen offen ansprechen und miteinander diskutieren. Die Teilnahme mehrerer Personen am Entscheidungsprozeß wirkt bei nicht optimal strukturiertem Vorgehen der Tendenz zur Über- oder Unterbetonung bestimmter Merkmale entgegen und schafft so eine sachlichere Grundlage für die zu treffende Personalentscheidung.

Die Planung, Durchführung und Auswertung von Eignungsinterviews ist eine außerordentlich komplexe Tätigkeit, die nur nach entsprechenden theoretischer Einführung und praktischer Einübung und mit sich immer wiederholendem Individuellem Feedback professionell zu bewältigen ist. Ein intensives Training sowie Feedback zur Qualität der Interviewtätigkeit sind daher unverzichtbar. Eine Möglichkeit zur Selbstdiagnose bzw. Selbstfeedback bietet das Diagnoseinstrument zur Erfassung der Interviewerkompetenz in der Personalauswahl (DIPA) (Strobel & Westhoff, 2002). Es gestattet sowohl Fremd- wie auch Selbstdiagnose von Interviewkompetenzen in der Personalauswahl bei der Planung, Durchführung und Auswertung von Eignungsinterviews. Durch wiederholte systematische Befragungen zahlreicher Experten (Delphistudie) wurden Richtwerte (Benchmarks) für konkrete Interviewerverhaltensweisen erhoben.

Die Güte des Eignungsinterviews

Im Bezug auf die Gütekriterien stellt das Interview einen Sonderfall da. Die Objektivität (intersubjektive Übereinstimmung) und die Paralleltest-Reliabilität fallen zusammen, da der Interviewer dem „Messinstrument“ entspricht. In Folge kann hier von Inter-Rater-Reliabilität gesprochen werden. Die Objektivität wird im Sinne der Inter-Rater-Reliabilität als Maß der Messgenauigkeit genutzt und liegt laut aktuellen Sammelreferaten im Mittel um 0.7 und höher (Strobel & Westhoff, 2004).

Rechtliche und ethische Rahmenbedingungen

Grundsätzlich dürfen eignungsdiagnostische Untersuchungen nur dann durchgeführt werden, wenn ein sachlicher Grund vorliegt, d.h. wenn überprüft werden soll, inwieweit eine Person den Tätigkeitsanforderungen einer Stelle entspricht. Es ist zu beachten, dass mit der Verwendung eines Interviews zur Eignungsbeurteilung in den Persönlichkeitsbereich des Interviewten eingegriffen wird. Es dürfen daher nur solche Merkmale des Interviewten erfasst werden, die für die vorgesehene Tätigkeit erforderlich sind und nicht oder nicht hinreichend durch Informationen aus anderen Informationsquellen erkennbar sind (Westhoff & Strobel, 2004).

Dies verlangt eine fundierte Anforderungsanalyse, da erfolgsrelevante Merkmale erst ermittelt werden müssen, um das Interview auf sie abzustimmen. Für die Zulässigkeit von

Fragen können die Hinweise zur Gestaltung von Personalfragebögen herangezogen werden (Schuler, 2002).

Zu unzulässigen Fragebereichen gehören:

1. Fragen zur Familie, sofern sie Heiratsabsichten oder intime Beziehungen betreffen. Ausnahmen stellen Fragen zu Geburtsdaten von Ehepartner und Kindern dar. Bei geplanten Auslandeinsätzen darf auch nach dem Beruf des Ehepartners gefragt werden.
2. Fragen zum Gesundheitszustand im Hinblick auf allgemeine Informationen zu früheren und derzeitigen Erkrankungen. Der Interviewte ist jedoch verpflichtet, über ansteckende Krankheiten zu informieren, ebenso über Einschränkungen, die seine Leistung am Arbeitsplatz massiv beeinträchtigen würden.
3. Fragen zu Vermögensverhältnissen oder zum letzten Einkommen, falls dies keinen Bezug zu Eignung oder Position hat. Ausnahmen stellen hier Bewerbungen als leitende Angestellte oder in Vertrauensstellungen dar.
4. Fragen zu Vorstrafen bei mangelnder Einschlägigkeit. Einschlägigkeit ist zum Beispiel die Frage nach Unterschlagungen bei der Bewerbung auf die Stelle als Kassierer.
5. Fragen zu Religions- oder Parteizugehörigkeit oder auch Gewerkschaftszugehörigkeit sind mit sehr wenigen Ausnahmen unzulässig. Ausnahmen sind hier Bewerbungen in so genannten Tendenzbetrieben wie beispielsweise konfessionelle Einrichtungen.
6. Fragen zu einer möglichen Schwangerschaft und können nur in begründeten Einzelfällen gestellt werden, d.h. wenn durch die Tätigkeit eine Gefährdung für Mutter oder Kind bestehen könnte.

Werden unzulässige Fragen gestellt, so darf der Interviewte sie wahrheitswidrig beantworten, ohne dass dem Arbeitgeber daraus später ein Anfechtungsgrund für den Arbeitsvertrag entstünde. Diese Möglichkeit bringt den Interviewten trotz allem in eine Konfliktsituation, da er/sie entweder wissentlich lügt und das Vertrauensverhältnis zum Arbeitgeber schon zu Beginn einschränkt oder aber mit einer wahr gemäßen Antwort unter Umständen seine/ihre Einstellungschancen mindert.

Wie für jedes eignungsdiagnostische Verfahren gilt auch für das Interview, dass es wissenschaftlichen Ansprüchen genügen muss und grundlegende Regeln beachtet werden sollten. Für das Interview sind dies die folgenden (Strobel & Westhoff, 2004):

1. Anforderungsorientierung: Jeder Interviewer sollte Fragen stellen, die in ihrer Zielrichtung an einer Anforderung des vorher erstellten Anforderungsprofils ausgerichtet sind, und das zu beurteilende Verhalten erfassen.
2. Verhaltensorientierung: Die im Interview gestellten Fragen sollen konkretes Verhalten in konkreten Situationen erfassen.

3. Transparenz: Jeder Gesprächsabschnitt im Interview sollte die Erläuterung beinhalten, nach welchem Verhalten bzw. Welcher Anforderung gefragt wird. Dabei sollte zum einen deutlich werden, welchen Bezug das erfragte Verhalten zur zu besetzenden Position hat, zum anderen auch, wie das Verhalten bzw. die Anforderung definiert sind.
4. Kooperative Grundhaltung zwischen Interviewer und Interviewten: Dies bedeutet, dass die Interviewsituation als gemeinsame Problemlösung anzugehen ist, wobei weder strategische Methoden des „Hervorlockens“ von Antworten zu nutzen sind, noch Druck auf den Interviewten ausgeübt werden sollte.
5. Respektieren von Persönlichkeit und Würde: Die Persönlichkeit und Würde des Interviewten sich auch in der Art der gestellten Fragen zu respektieren, d.h. beispielsweise ohne Herabsetzen des Interviewten durch die Formulierung der Frage.

Aufzeichnungen aus dem Eignungsinterview unterliegen dem Datenschutz. Die Erfassung, Speicherung und Verwendung der Daten dürfen nur nach vorheriger Information der untersuchten Person und ihrem abgegebenen Einverständnis erfolgen. Des Weiteren darf die Verwendung lediglich zweckgerichtet erfolgen, d.h. grundsätzlich dürfen die erhobenen Informationen nur für den Zweck verwendet werden, der der untersuchten Person bekannt gegeben war und dem sie zugestimmt hat. Sollten die Daten für andere Zwecke als die berufsbezogene Eignungsbeurteilung genutzt werden, dann darf dies auch wieder nur nach vorhergehender ausführlicher und expliziter Information der untersuchten Person sowie ihrem eingeholten Einverständnis erfolgen.

Zusammenfassung

Bewertungen der Passung von Person und Organisation kann auf mehrere eignungsdiagnostische Verfahren zurückgreifen. Durch die Verwendung multipler Verfahren kann durch sequentielle Informationsanreicherung, die Passungsentscheidung optimiert werden. In dem obigen Beitrag wurden einige wichtige Punkte zur Planung, Durchführung und Auswertung des Eignungsdiagnostischen Interviews zusammengefasst. Durch eine Systematisierung der Vorgehensweise des Eignungsinterview, besteht die Möglichkeit eine hochwertigere Informationsbasis für Auswahlentscheidungen zu sichern.

Literatur

Campion, M.A., Palmer, D.K. & Campion, J.E. (1997). A review of structure in the selection interview. *Personnel Psychology, 50,* 655-702.

Frieling, E. & Hoyos, C.G. (1978). *Fragebogen zur Arbeitsanalyse (FAA): Deutsche Bearbeitung des Position Analysis Questionaire (PAQ)*. Bern: Huber.

Janz, T., Hellervik, L. & Gilmore, D.C. (1986). *Behavior description interviewing.* Boston: Allyn & Bacon.

Latham, G.P., Saari, L.M., Pursell, E.D. & Campion, M.A. (1980). The situational interview. *Journal of Applied Psychology, 17*, 422-427.

Mael, F.A. (1991). A conceptual rationale for the domain and attributes of biodata items. *Personnel Psychology, 44,* 763 – 792.

Reimann, G. (2004). Arbeits- und Anforderungsanalyse. In Westhoff, K., Hellfritsch, L. J., Hornke, L. F., Kubinger, K. D., Lang, F., Moosbrugger, H., Püschel, A., & Reimann, G. (Hrsg.). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430*. Lengerich: Pabst Science Publishers.

Schuler, H. (1992). Das Multimodale Interview. *Diagnostica, 38,* 281 - 300.

Schuler, H. (1998). Noten und Studien- und Berufserfolg. In Rost, D.H. (Hrsg.): *Handwörterbuch Pädagogische Psychologie* (S. 370-374). Weinheim: Psychologie Verlags Union.

Schuler, H. (2000). *Psychologische Personalauswahl.* Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. (2002). *Das Einstellungsinterview*. Göttingen: Hogrefe.

Schuler, H. & Marcus, B. (2001). Biographieorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 176-212). Göttingen: Hogrefe.

Strobel, A. & Westhoff, K. (2002). Entwicklung eines Diagnoseinstruments zur Erfassung von Interviewkompetenz in der Personalauswahl (DIPA): Grundlagen und Konstruktion. Untersuchungen des Psychologischen Dienstes der Bundeswehr, Band 36/3, 83-132.

Westhoff, K. & Kluck, M.-L. (2003). *Psychologische Gutachten schreiben und beurteilen*. (Vierte, vollständig überarbeitete und erweiterte Auflage). Berlin: Springer.

Westhoff, K. & Strobel, A. (2004). Eignungsinterview. In Westhoff, K., Hellfritsch, L. J., Hornke, L. F., Kubinger, K. D., Lang, F., Moosbrugger, H., Püschel, A., & Reimann, G. (Hrsg.). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430*. Lengerich: Pabst Science Publishers.

Westhoff, K. (2004). Konstrukte und Operationalisierungen. In Westhoff, K., Hellfritsch, L. J., Hornke, L. F., Kubinger, K. D., Lang, F., Moosbrugger, H., Püschel, A., & Reimann, G. (Hrsg.). *Grundwissen für die berufsbezogene Eignungsbeurteilung nach DIN 33430*. Lengerich: Pabst Science Publishers.

Self-Assessment als Studienberatung und Bewerbervorselektion

Annette Höpfner

Fragestellung

Im Rahmen zukünftiger verstärkter Studienbewerberauswahl durch die Hochschulen stellt sich die Frage, inwieweit ein Self-Assessment im Internet zur Auswahl und Beratung der Interessenten nützlich sein kann. Grundlage ist die Feststellung hoher Studienabbrüche, häufiger Studienwechsel und langer Studiendauer, für welche einige Gründe bei den Studenten selbst gesehen werden wie z.B. falsche Erwartungen oder mangelhafte Studierfähigkeit andere in der Umwelt zu finden sind wie Studienbedingungen, Curricula etc.

Es ist zunächst auszuloten, welches konkrete Ziel mit dem Self-Assessment verfolgt werden soll, Studienerfolg oder die Selektion der Bewerber. Die Prognose des Studienerfolgs über die Schulleistungen und die Ergebnisse kognitiver Tests analog zu den in der Arbeitswelt eingesetzten Einzelmethoden ist vielversprechend. Psychometrische Testverfahren haben sich bereits bei den Medizinern als Erfolg versprechend für die Entscheidung für oder gegen ein Studium erwiesen. PC-basierte Verfahren sind in der Lage, ökonomisch und interaktiv verschiedene Tests und Aufgabenstellungen zu integrieren.

Ein webbasiertes Self-Assessment eröffnet wie wir sehen werden eine zeitgemäße Möglichkeit zum kombinierten Einsatz konstruktorientierter Verfahren kognitiver und nicht-kognitiver Parameter, die bisher nur kostenintensiv in der Verbindung von persönlichen Interviews mit parametrischen Tests erhoben wurden.

Analog zum Job-Fit-Ansatz (PJF, Amelang, 1997) der Arbeits- und Organisationspsychologie ist das Ziel diejenigen Studenten zu identifizieren, die eine hohe Passung zur Universität bzw. dem Studiengang aufweisen. Andererseits kann durch die Rückmeldung der Auswertung der Ergebnisse dem Studieninteressenten wertvolle Information zur Entscheidungsfindung an die Hand gegeben werden. Angehende Studenten sind mit dem Internet bereits sehr vertraut und informieren sich anhand dieses Mediums häufig über die Universitäten und ihre persönlichen beruflichen Möglichkeiten.

Zur Entscheidung für oder gegen ein Studium tragen oft Aussagen über die Arbeitsmarktperspektive bei. Über ein solches Tool erhält der Interessent weitergehende Informationen, die ihm helfen, solche Aussagen zu relativieren und mehr nach seinen eigenen Interessen und Fähigkeiten zu entscheiden (intrinsische Motivation) als nach der vermeintlich aktuellen beruflichen Modeerscheinung (extrinsische Motivation).

In dieser Arbeit möchte ich zunächst neben einigen methodischen Anmerkungen auf die Studienzufriedenheit und den Studienabbruch eingehen, danach den Auftrag der Studienberatung und Studentenauswahl erläutern und anschließend vorstellen, wie die Universität Aachen ihr Online-Self-Assessment aufgebaut hat und in ihre Beratung einbezieht.

Metaanalytische Befundlage und Intelligenz

Nach der Zusammenfassung der metaanalytischen Befundlage nach Schmidt und Hunter (1998, zitiert nach Schuler, 2001) sind die berufseignungsdiagnostischen Verfahren mit der höchsten Validität die allgemeinen kognitiven Fähigkeitstests (.51), das strukturierte Einstellungsgespräch (.51) und Fachkenntnistests (.48).

Vergleicht man allgemeine mit spezifischen kognitiven Fähigkeitstests, so sprechen zunächst zwei Argumente für den allgemeinen Test von Fähigkeiten: er ist kostengünstiger, da generelle Fähigkeiten sich wenig wandeln und die Tests somit nicht so häufig überarbeitet werden müssen, außerdem lassen sich für die Bestimmung seiner Validität in der betrieblichen Praxis schneller große Stichproben finden. Je nach Sparte schwer miteinander vergleichbare Spezialverfahren sind sehr kostenaufwendig und müssen hinsichtlich der Aktualität ihrer verwendeten Komponenten regelmäßig überprüft werden. Für den universitären Bereich kann man rasch mit großen Stichproben rechnen, sodass es sich bei angemessenem Aufwand für die Spezialbereiche einzelner Fachbereiche bzw. Profile einzelner Universitäten durchaus lohnen könnte, sowohl allgemeine als auch spezifische Testverfahren zum Einsatz zu bringen. Dies gilt vor allem vor dem Hintergrund der wahrscheinlichen Vorselektion nach kognitiven Fähigkeiten bei Studienbewerbern. Es ist zu prüfen, inwieweit die hohe Validität auch hier erreicht wird und die inkrementelle Validität der speziellen Verfahren nicht deshalb höher ausfällt. Auch lässt sich die Prognosequalität womöglich durch den Blick auf weitere Einflussfaktoren erhöhen und mehr Gewissheit über Ursachenzusammenhänge und damit höhere Konstruktqualität erreichen.

Einstellungsgespräche sind im universitären Betrieb personell nicht abdeckbar. Es besteht aber die Möglichkeit, die jeweiligen Themen über entsprechend integrierte Spezialtests abzudecken, was hinsichtlich der Validität und Reliabilität ehedem zu bevorzugen ist.

Analog zur Empfehlung, im betrieblichen Bereich Fachkenntnistests einzusetzen, die der Arbeitstätigkeit möglichst ähnlich sind, kann dies auch für studiumsrelevante Kenntnisse umgesetzt werden. Dies setzt aber eine sorgfältige Anforderungsanalyse des jeweiligen Faches voraus, da ansonsten keine ausreichende Validität zu erwarten ist.

Intelligenz gilt als die Basis für Fachwissen, was man direkt auf das Studium übertragen kann. Sie stellt das Potenzial zum ganz allgemeinen Erwerb von Kenntnissen bereit.

Bei der Erstellung von eignungsdiagnostischen Verfahren für die Studienbewerberauswahl sollten ebenso wie im betrieblichen Bereich die Kriterien der Testbeurteilung angewandt

werden, um ethischen und rechtlichen Aspekten mit wissenschaftlicher Professionalität Rechnung zu tragen.

Ein weiterer Aspekt ist, dass laut Amelang (1997) selbst- und fremdbeurteilte Leitungsvoraussetzungen, leistungsfördernde Arbeitshaltung sowie die Rolle der im universitären Umfeld erfahrenen Anerkennung und Integration einen wichtigen Einfluss auf das Kriterium zu haben scheinen. Überdies korrelieren diese Maße nicht mit dem typischerweise herangezogenen Kriterium der Schulabschlussnote. Daraus folgert er eine komplexere Operationalisierung für „Studierfähigkeit".

Problemfeld: konstruktorientierte Verfahren

Bei der Verfahrenskombination, wie sie bei einem Assessment erfolgt, wird davon ausgegangen, dass die Kombination mehrerer Verfahren bessere Vorhersagen liefert als die isolierten Einzelkomponenten. Vorher als hoch kriteriumsvalide nachgewiesene Verfahren können sich dabei jedoch plötzlich als redundant herausstellen oder einem Moderatoreffekt unterliegen.

Schmidt und Hunter (1998) empfehlen, zunächst die Verfahren zu Erhebung der allgemeinen kognitiven Fähigkeiten einzusetzen und weitere nach ihrem inkrementellen Zuwachs hinzuzunehmen. Dabei ist wiederum der Rückgriff auf die Ergebnisse einer sorgfältigen Arbeits- und Anforderungsanalyse von großem Vorteil.

Fachspezifische vs. fachübergreifende Prädiktion

Bei der fachspezifischen Prädiktion ergeben sich nach Amelang (1997) drei Dimensionen:

1. mathematisch-naturwissenschaftlich vs. kultur- und geisteswissenschaftlich.
2. Differenzierung des Leistungsniveaus nach Schulnoten, intellektuellen Fähigkeiten und Selbsteinschätzung.
3. soziale und pädagogische Neigungen vs. Interesse an Politik und Wirtschaft

Über die Stellung der Psychologie hinsichtlich dieser drei Dimensionen findet sich bei Amelang keine Aussage.

Bei Untersuchungen fand sich, dass die Prognose z.B. von Prüfungsnoten genauer wird, wenn pro Studienbereich eine Prognosegleichung verwendet wird. Hingegen ließ sich die Studienzufriedenheit besser mit einer fachübergreifenden Prognose vorhersagen. Prognosesysteme sind also durchaus für größere Studienfelder geeignet, wenn diese ähnliche kognitive oder motivationale Anforderungen stellen. Die Grenzen lassen sich durch empirische Forschung ermitteln. In jedem Fall ist unter Einbeziehung zusätzlicher Verfahren die Profilbildung für einzelne Hochschulen bzw. die fachspezifische Profilbildung zweckmäßig.

Interesse und Motivation

Nach Schuler (2000) sind Interessen vor allen anderen Parametern die Bestimmungsgrößen der Selbstselektion. Er hält es für plausibel, dass sie daher auch als Prädiktoren der Berufszufriedenheit, Fluktuation und ähnlicher Kriterien nutzbar sind. Dies ließe sich wahrscheinlich auch auf das Studium übertragen.

Zur Messung von Leistung und Motivation haben sich Fragebögen bewährt. Es gilt zwischen relativem und generellem Interesse zu unterscheiden. Relatives Interesse wird im betrieblichen Kontext über die Rangzuordnung fachspezifischer Aussagen über Situationen im Vergleich gemessen. Für generelles Interesse werden unterschiedliche Interessensbereiche ins Verhältnis gesetzt, die keinen direkten Bezug zu bestimmten Fachrichtungen aufweisen. Holland nimmt für seinen Allgemeinen Interessen-Struktur-Test (deutsche Fassung von Bergmann & Eder, 1992) an, dass Interesse eine grundlegende Persönlichkeitsorientierung ist und damit die Berufswahl als durch allgemeine Wesensmerkmale der Person beeinflusst. Holland analysiert auf Basis dieser Theorie und mit Hilfe einer Typologisierung nach Interessenlage die Kongruenz zwischen den Orientierungen der Person und den Anforderungen der Umwelt.

Der Person-Job-Fit-Ansatz (PJF, Amelang, 1997) der Personalauswahl verwendet ebenso diese hier Passung genannte Kongruenz.

Verfälschungstendenzen

Verfälschungstendenzen wie die Beantwortung von Fragen nach dem Kriterium der sozialen Erwünschtheit stellen heute weniger Probleme dar als früher angenommen. Hat man noch vor einiger Zeit mit sogenannten Lügenskalen versucht, diese Tendenzen zu kontrollieren, so wird dies heute kaum noch auf diese Weise behandelt.

Schuler (2000) weist darauf hin, dass es wichtiger ist, inwieweit durch Verfälschungstendenzen die Validität der Untersuchung in Gefahr ist, als dass in Auswahlsituationen beschönigende Selbstdarstellung erfolgt. Er legt die Annahme nahe, dass die Neigung zu positiver Selbstdarstellung zunächst nicht ausschließlich zur Verfälschung anderer Persönlichkeitsmerkmale führt, sondern teilweise zur Substanz dieser Merkmale gehört. Die Validität von Persönlichkeitsmaßen werde also durch die Selbstdarstellung nicht beeinträchtigt.

Andererseits kann davon ausgegangen werden, dass der Bewerber erwartet, später an seinen Angaben gemessen zu werden. Im Kontext des Studiums würde er das an seinen Leistungen im Verhältnis zu denen anderer Studierender bemerken, was bei schlechtem Abschneiden für wenig Anerkennung durch die Kommilitonen und Lehrkräfte sorgen würde. Dies dürfte dazu beitragen, dass die Verfälschungstendenz zum Zwecke eines guten Eindrucks in Grenzen gehalten wird.

Akzeptanz

Verwender von Testinstrumenten und Auswahlverfahren sowie Diagnostiker vernachlässigen in der Regel die Sichtweise der Bewerber. Kritiker warnen vor allem vor Intransparenz und fraglicher Tätigkeitsrelevanz (Schuler, 2000). Ein Einstellungsangebot anzunehmen hängt aber ganz entscheidend von dem Eindruck, den ein Interviewer beim Bewerber macht ab (Schmitt und Coyle, 1976). Der Interviewer ist Repräsentant der Institution. Dies ließe sich auch auf Hochschulen übertragen. In gewissem Sinne sind der Internetauftritt und die Möglichkeit der Informationsbeschaffung über eine Hochschule ebenfalls imagefördernd oder -schädigend. Informationsbeschaffung und Bewerbungsentscheidung sind ein interaktiver Prozess.

Schuler und Stehle (1983) schlagen vier Parameter des Erlebens von Auswahlsituationen vor, die sie insgesamt als „soziale Validität“ bezeichnen.

Ist die soziale Validität nicht gegeben, bewerben sich erstens nicht diejenigen Studenten bei der Hochschule, welche die Hochschule sich für wünscht und zweitens steht und fällt das Prognosepotential der verwendeten Tests damit.

Soziale Validität

Die soziale Validität hängt von den Parametern Information, Partizipation, Transparenz und Urteilskommunikation ab. Zu diesem Konzept wurden eine Reihe von Untersuchungen durchgeführt. Als Zwischenresümee ergibt sich beispielsweise, dass Fähigkeits- und Leistungstests besser akzeptiert werden als Persönlichkeitstests sowie, dass mündlich erhaltene Informationen bevorzugt und für glaubwürdiger eingestuft werden (Schuler, 1990). Die vorläufigen Ergebnisse sowie die für diese Untersuchungen entwickelten Fragen zu allen vier Dimensionen könnten sowohl für die Überlegungen zur Konzeption eines Self-Assessment nützlich sein als auch in das Instrument selbst integriert werden.

Organisation

Aus der Organisationspsychologie abgeleitete Betrachtungen zu Eigenschaften von Organisationen können gerade vor dem Hintergrund des steigenden Kosten- und Effizienzdrucks auf universitäre Einrichtungen von Interesse sein. Wenngleich sich der Bezug zur Personalauswahl im betrieblichen Bereich noch wenig stringent nachweisen lässt, so sind diese Überlegungen zumindest von plausibler Bedeutung für Strategie und Gewinnung von Studierenden und wissenschaftlichen Mitarbeitern. Im Rahmen der fortschreitenden Globalisierung und der internationalen Anpassung von Studienabschlüssen wird dies besonders deutlich.

Prozesse (z.B. Interaktionen) sind geprägt von den Organisationszielen. Dabei geht es wesentlich darum, welche Art von Leistung gefordert wird und welche Art von Belohnungen dafür zur Verfügung steht. Ist für einen Studierenden über den reinen

Scheinerwerb hinaus keine weitere Belohnung erkennbar, lässt sich durchaus die Frage nach seiner Bindung, seinem Engagement und seiner Haltung zu der von ihm besuchten Hochschule stellen. Individuelles Leistungsverhalten ist geleitet von Interessen, Bedürfnissen und Werthaltungen. Auch die Organisation als Ganzes kann ihre Prozesse besser strukturieren je klarer jedem Mitglied die Ziele sind. Weiterhin könnten Überlegungen zu Organisationsstil und Organisationskultur an Hochschulen fruchtbare Diskussionen entfachen. Wie ist Hierachie, Autorität und Verantwortung aufgebaut? Wie beeinflussen Werte und Normen die Wahrnehmung und das Verhalten der Organisationsmitglieder? Hier stößt man unweigerlich wieder auf die Passung zwischen Person und Umwelt. Stark ausgebildete Organisationskulturen stellen sowohl eine Quelle des Erfolges dar als auch die Gefahr der Überhomogenisierung. Dies gilt es dabei beides zu bedenken (Schuler, 2001).

Studienzufriedenheit

Nach der Person-Environment-Fit-Theorie von Caplan (P-E-Fit-Theorie, 1987) ist die Übereinstimmung von Person und Arbeitsumwelt als ein entscheidender Faktor für Stress verantwortlich. Die Umwelt stellt Anforderungen an die Person, der diese mit Fähigkeiten zur Bewältigung der Anforderungen begegnet. Die Person hat Bedürfnisse, für welche die Umwelt Angebote bereit hält. Stress tritt dann auf, wenn die Fähigkeiten der Person nicht den Anforderungen entsprechen und die Angebote der Umwelt nicht den Bedürfnissen der Person. Die Entstehung von Stress ist in diesem Fall von der subjektiven Wahrnehmung gesteuert nicht von den objektiven Gegebenheiten. Wichtig ist es also, wie die Person die Passung beurteilt.

Aus dieser Theorie lässt sich prognostizieren, dass die allgemeine Studienzufriedenheit der Studierenden erstens umso größer ist, je besser ihre Fähigkeiten mit den im Studium gestellten Anforderungen übereinstimmen. Zweitens determiniert das Ausmaß, in dem das Angebot durch die Hochschule und das Studium den Bedürfnissen der Studierenden entspricht, die Zufriedenheit (Heise, Westermann, Spiess & Stephan, 1997).

Heise et.al. haben in ihren Untersuchungen fünf Faktoren extrahiert, welche die sogenannten Fit-Dimensionen ausmachen.

1. Psychische und kognitive Bewältigungsfähigkeiten	Fähigkeiten und Anforderungen
2. Lernstrategien	
3. naturwissenschaftlich-mathematische Begabung	
4. Berufliche Orientierungen nach Holland	Bedürfnisse und Angebote
5. Fachspezifisches Studieninteresse	

Dabei wurden Mediziner, Psychologen und Juristen auf drei Skalen untersucht:

1. Zufriedenheit mit den Studieninhalten
2. Zufriedenheit mit den Studienbedingungen
3. Zufriedenheit mit der Bewältigung von Studienbelastungen

Auffällig war, dass bei Studierenden der Psychologie auf der Dimension Fähigkeiten-Anforderungen bei der Zufriedenheit mit der Bewältigung der Studienbelastungen der Faktor naturwissenschaftlich-mathematische Begabung die größte Rolle spielte. Bei der Zufriedenheit mit den Studienbedingungen war in dieser Stichprobe der Fit zwischen Bedürfnissen und Angeboten wichtig im Gegensatz zur Rechtswissenschaft, in der sich der Fit zwischen Anforderungen und Fähigkeiten als maßgeblich herausstellte.

Insgesamt lässt sich festhalten, dass die P-E-Fit-Dimensionen für die Prädiktion der allgemeinen Studienzufriedenheit verwendbar sind. Die Bedeutung der einzelnen Faktoren hängt davon ab, welcher Aspekt der Studienzufriedenheit betrachtet wird. Das fachspezifische Studieninteresse ist für die Vorhersage der Zufriedenheit mit den Studieninhalten am bedeutsamsten.

Studienabbruch

Um Studienabbrüche zu untersuchen kann man entweder die Abbrecher retrospektiv befragen oder prospektiv Kausalmodelle zur Erklärung des Abbruchverhaltens spezifizieren. Bei der retrospektiven Befragung ergibt sich die Problematik einer eventuellen Verzerrung durch Antworttendenzen und einen möglichen Selektionsprozess, dem Erinnerungen unterliegen sowie Orientierung an individuellen Attributionsstilen. Deshalb dient die direkte Befragung in ihrer heuristischen Funktion hauptsächlich der Hypothesengenerierung.

In umfangreichen älteren und neueren Studien weisen Studienabbrecher im Wesentlichen folgende Merkmale auf (Gold, A. und Kloft, C., 1991):

1. ausgeprägtere Leistungs- und Motivationsprobleme
2. emotionale Schwierigkeiten
3. Orientierungsprobleme im universitären Umfeld
4. nichtakademische Berufsperspektiven
5. Konflikte zwischen Berufsausbildung und Familie
6. unrealistische Studienerwartung

Allein die vorangegangenen Studienleistungen sowie das Ausmaß an Studienzufriedenheit hatten einen direkten Einfluss auf die Abbruchentscheidung bei Studierenden ingenieurwissenschaftlicher Fächer an der TH Aachen. Merkmale des Arbeitsverhaltens

finden indirekt über zwischengeschaltete Prüfungserfolge ihren Niederschlag (Ströhlein, 1982, zitiert nach ebd.).

Gold und Kloft fanden 1991 fachübergreifende Abbruchsquoten von mehr als 14%. Die höchsten Quoten wiesen die Recht- und Wirtschaftswissenschaften auf (Abbruch relativ spät im Verlauf des Studiums), die niedrigsten die Naturwissenschaften (Abbruch kurz nach Beginn des Studiums). Retrospektiv nannten die Befragten häufig schlechte Berufsaussichten, Interessenänderungen, Lernschwierigkeiten und den Wunsch nach einer praktischen Tätigkeit. Beklagt wurde von 80% die Praxisferne des Studiums und attraktiv erscheinende nichtakademische Berufsalternativen. Externe Belastungen finanzieller, familiärer oder gesundheitlicher Art fanden sich weniger häufig.

Lernprobleme scheinen ein zentrales Thema beim Studienabbruch darzustellen. Leistungsschwierigkeiten, eine unzureichende Motivation und fehlende Anerkennung durch Kommilitonen sind entscheidende Prädiktoren der prospektiven Analyse. Männer sind von Lernproblemen offenbar häufiger betroffen als Frauen. Prognostisch relevant sind die Aspekte der selbst- bzw. fremdbeurteilten Leistungsvoraussetzungen, die Bedeutsamkeit leistungsfördernder Arbeitshaltung sowie die Rolle der im universitären Umfeld erfahrenen Anerkennung. Persönlichkeitsmerkmale im engeren Sinne leisten keinen entscheidenden Beitrag zur Differenzierung zwischen Abbrechern und Nichtabbrechern, nur zu einer „Vulnerabilität“. Schuler (2000) berichtet von - mangels für eine umfassende Metastudie notwendiger Primärstudien - bisher wenig aufgeklärten Moderatoreffekten hinsichtlich der Leistungsprognose anhand der Big Five.

Individualprognosen sind also allein anhand psychologischer Bedingungsvariablen kaum zuverlässig möglich. Beratungs- und Interventionsmaßnahmen sollten daher die subjektive Sichtweise der Betroffenen einbeziehen. Arbeitsmarktpolitische Missstände, defizitäre Studienbedingungen und belastende Lebensumstände dürfen bei der Betrachtung ebenso nicht außer Acht gelassen werden.

Studienberatung

Studienberatung soll zur nachhaltigen und richtigen Entscheidungsfindung beitragen, beschränkt sich aber derzeit nicht zuletzt wegen knapper Personalressourcen auf organisatorische Aspekte der Bewerbung um einen Studienplatz und einige wenige inhaltliche Informationen zu den Fachrichtungen allgemein. In der Regel ist die Inanspruchnahme von Studienberatung für den Bewerber ein unangenehmes Unterfangen, weil es mit langen Wartezeiten, Unsicherheit der Zuständigkeit und gegebenenfalls weiter und damit kostenintensiver Anreise an die interessierende Hochschule verbunden ist.

Der Wissenschaftsrat (2004) empfiehlt eine Professionalisierung der Studienberatung. Bei gleichzeitigem Kostendruck dürfte sich dies eher nicht im Ausbau der Personalplanung für etwaige hauptberufliche Studienberater oder gar wie vom Rat gefordert in der Schaffung

eines speziellen Berufsbildes und -ausbildung zum solchen realisieren lassen. Gleichwohl ist einzusehen, dass die Profilbildung der Hochschulen gerade im internationalen Vergleich von hoher Bedeutung ist und Studienfächer sich immer weiter ausdifferenzieren. Beidem muss in irgendeiner Form Rechnung getragen werden. Die gegenwärtige Beratung und die praktizierten Auswahlverfahren dienen selten dazu, den Grad der Motivation für ein bestimmtes Fach oder die (Fehl-)Vorstellungen darüber festzustellen.

In einem kombinierten webbasierten und damit interaktiven Testverfahren, das sich neben den allgemeinen kognitiven und nicht-kognitiven Aspekten auch an Fachspezifika festmacht, kann Information über das Studium und die Anforderungen und Gegebenheiten der jeweiligen Fachrichtung implizit erfolgen. Traditionell blieb dies den persönlichen Interviews vorbehalten.

Gerade bei Psychologen sollten zum Beispiel interaktive Kompetenzen in der Auswahl berücksichtigt werden (Rindermann, H., und Oubaid, V., 1999).

Online-Self-Assessment am Beispiel der TH Aachen

Als Self-Assessment wird die berufsbezogene, eignungsdiagnostische Messung eines Konstruktes über verschiedene Messmethoden hinweg definiert (in Anlehnung an Schuler, 2001), das ein Bewerber selbständig durchführen kann.

Generell hat die computergestützte Eignungsdiagnostik auch ohne das Internet einige Vorteile. Dies sind unter anderem die volle Standardisierung und Rationalisierung der Durchführung und Auswertung sowie Kontrolle der Zeitvorgabe für Items und Verfahrensteile, schneller Zugriff auf Ergebnisse, erhöhte Kontrolle des Urheberrechtsschutzes, reduziertes Impression Management der Bewerber und hohe Akzeptanz bei den Nutzern (Schuler, 2001).

Im Rahmen von mehreren Diplomarbeiten ist das erste deutsche, webbasierte, frei zugängliche Self-Assessment für die Studienbewerberauswahl an der TH Aachen für die Fächer Informatik, Elektrotechnik und Technische Informatik entstanden. Es handelt sich um ein in seiner grafischen Gestaltung zielgruppenadäquates Tool von hoher technischer Performanz. Ein gewisser Unterhaltungswert ist gegeben und wird durch eine die Testperson begleitende marsmännchenähnliche Comicfigur unterstrichen.

Ziel

Mit dem Aachener Self-Assessment soll ein psychometrisch fundierter, ökonomischer Test mit möglichst hoher Reliabilität und Validität entwickelt werden (Zimmerhofer, 2004a). Minderung der Studienabbruchsquote und Steigerung des Studienerfolgs über verstärkte Selbstreflexion und mehr Information über das Studium und die Hochschule sind die primären Ziele. Die TH Aachen verspricht sich durch dieses Tool eine Optimierung der Passung zwischen Studierenden und ihrer Studienentscheidung. Gedruckte Informationen

sind zwar immer noch gefragt, das Medium Internet hat aber inzwischen aufgeholt und eine sehr hohe Akzeptanz und Nutzung bei der Zielgruppe erreicht nicht nur bei Interessenten der Informatik. Die aktuelle Version im Netz befindet sich in der Testphase. Erste aussagekräftige Ergebnisse zur Validität liegen vor und werden im Rahmen einer Dissertation aufgegriffen und der Ansatz weiterentwickelt. Täglich gehen zusätzliche Datensätze ein. Es wird ganz explizit dazu aufgerufen, dass sich auch Nicht-Studenten, bereits Studierende und auch Studieninteressenten anderer Fächer testen, um die Datenbasis zu erhöhen und später über einen Extremgruppenvergleich Diskriminanzanalysen durchzuführen. Weitere Expertengespräche sind geplant, Testungen unter Aufsicht sollen hinzugezogen werden. Über die Verwendung von Umweltnormwerten, die den Vergleich zu Gleichaltrigen ermöglicht, soll die Korrelation zwischen Kongruenz von Umwelt (Universität und Studienfach) und Person (eigenes Interessenprofil) sowie den Studienleistungen erhöht werden (Bergmann, 1992)

Vorgehen

Für Interessenten erfolgt eine Anmeldung zum Self-Assessment im Internet über eine persönliche Emailadresse, an die ein automatisch erstelltes Passwort geschickt wird. Die Testperson kann sich danach mit ihrem Benutzernamen und dem Passwort in den Testbereich einloggen. Der Test wird ohne Unterbrechung durchgeführt. Pausen werden abschnittweise angeboten. Die Dauer beträgt ca. eine Stunde. Am Ende erhält der Testteilnehmer eine umfangreiche Rückmeldung. Der Test erfolgt anonym.

Struktur

Nach einer ausführlichen Anforderungsanalyse, bei welcher unter anderem die Critical-Incident-Technique nach Flanagan (CIT, 1954) verwendet sowie komplexe Interviews mit Studierenden geführt wurden, konnten die entscheidenden Dimensionen des Tests festgelegt werden. Inhaltlich ist das Self-Assessment folgendermaßen aufgebaut:

1. Explorative Daten inklusive persönlicher Angaben zu bisherigen Leistungen in Studium und Schule.
2. Intrinsische Lernmotivation
3. Extrinsische Lernmotivation
4. Selbstwirksamkeit
5. Handlungskontrolle
6. Logisches Denken
7. Mathematische Fähigkeiten
8. Matrizentest

9. Auswertung
10. gespeicherte jederzeit abrufbare Rückmeldung

Gegenüber der ersten Version wurde die Itemzahl um etwa 30% gemindert, sodass die Zeit für die Durchführung des Test nur noch etwa eine Stunde gegenüber vorher eineinhalb bis zwei beträgt. Dies war ohne größere Einbußen der Messgenauigkeit gemäß Cronbachs-Alpha möglich.

Auswertung und Rückmeldung

Am Ende des Tests wird automatisch eine ca. 10-seitige Rückmeldung generiert und dem Studieninteressenten zur Verfügung gestellt. Diese soll zum Nachdenken über das eigene Leistungs- und Interessenprofil anregen. Die Rückmeldung beinhaltet sowohl Erläuterungen zum Verfahren und den Rohwerten, Normwerten und Einschätzungen als auch anschauliche Grafiken. Über das persönliche Login kann die Testperson jederzeit auf ihre Ergebnisse und die Rückmeldung zugreifen. Der Testteilnehmer wird um sein Feedback und Verbesserungsvorschläge gebeten. Eine Verlosung von Büchergutscheinen dient als Dankeschön für die Teilnahme.

Literatur

Amelang, M., (1997). Differentielle Aspekte der Hochschulzulassung: Probleme, Befunde, Lösungen. In T. Herrmann (Hrsg.), *Hochschulentwicklung – Aufgaben und Chancen.* Heidelberg: Asanger.

Bergmann, C., (1992). Schulisch-berufliche Interessen als Determinanten der Studien- bzw. Studienwahl und –bewältigung. Eine Überprüfung des Modells von Holland. In A. Krapp & M. Prenzel (Hrsg.), *Interessen, Lernen, Leistung – Neuere Ansätze einer pädagogisch-psychologischen Interessensforschung* (S. 195-220). Münster: Aschendorf

Bergmann, C. & Eder, F. (1992). *Allgemeiner Interessen-Struktur-Test/Umwelt-Struktur-Test (AIST/UST).* Weinheim: Beltz.

Caplan, R.D. (1987). Person-Environment fit: Past, present, and future. In C.L. Cooper (Ed.), *Stress research: New directions for the 1980s* (pp. 35-78). London: Wiley.

Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin*, 51, 327-358.

Gold, A. & Kloft, C. (1991). Der Studienabbruch: Eine Analyse von Bedingungen und Begründungen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 2*, S. 265 –279.

Heise, E., Westermann, R., Spies, K. & Stephan, H. (1997). Die Übereinstimmung von Fähigkeiten und Bedürfnissen der Studierenden verschiedener Fächer mit den Anfor-

derungen und Angeboten im Studium als Determinante der Studienzufriedenheit. In Kittler, U., Metz-Göckel, H. (Hrsg.), *Pädagogische Psychologie in Erziehung und Organisation*. Essen: Verlag Die Blaue Eule.

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.

Schmidt, F.L. & Hunter, J.E. (1998). Messbare Personmerkmale: Stabilität, Variabilität und Validität zur Vorhersage zukünftiger Berufsleistung und berufsbezogenen Lernens. In M. Kleinmann & B. Strauss (Hrsg.), *Potentialfeststellung und Personalentwicklung* (S. 16-43). Göttingen: Hogrefe.

Schmitt, N. & Coyle, B.W. (1976). Applicant decisions in the employment interview. *Journal of Applied Psychology*, 61, 184-192.

Schuler, H. & Stehle, W. (1983). Neuere Entwicklungen des Assessment-Center-Ansatzes – beurteilt unter dem Aspekt der sozialen Validität. *Psychologie und Praxis. Zeitschrift für Arbeits- und Organisationspsychologie,* 27, 33-44.

Schuler, H. (1990). Personalauswahl aus der Sicht der Bewerber: Zum Erleben eignungsdiagnostischer Situationen. *Zeitschrift für Arbeits- und Organisationspsychologie, 34,* 184-191.

Schuler, H. (2000). *Psychologische Personalauswahl*. Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. & Höft, S. (2001). Konstruktorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 93 – 131). Göttingen: Hogrefe

Wissenschaftsrat (2004). Empfehlungen zur Reform des Hochschulzugangs. *Wissenschaftsrat*.

Zimmerhofer, A. (2003). *Neukonstruktion und erste Fassung eines webbasierten Selfassessments zur Feststellung der Studieneignung für die Fächer Elektrotechnik, Technische Informatik sowie Informatik an der RWTH Aachen.* Diplomarbeit. Lehrstuhl für Betriebs- und Organisationspsychologie. Institut für Psychologie der RWTH Aachen.

Zimmerhofer, A. (2004a). Weiterentwicklung eines webbasierten Studienberatungstests für die Fächer Informatik, Elektrotechnik und Technische Informatik an der RWTH Aachen. In Hornke, Lutz F., *Exposés / Ergebnisse der laufenden und abgeschlossenen Arbeiten.* Institut für Betriebs- und Organisationspsychologie der RWTH Aachen.

Zimmerhofer, A. (2004b). SelfAssessment. *HTTP://WWW.ASSESS.RWTH-AACHEN.DE* . 05.03.2005

Test für medizinische Studiengänge (TMS)

Stephanie Thomas

Einleitung

Bei 45.000 Bewerbungen pro Jahr und einem Angebot von 13.000 Studienplätzen wurde zum Wintersemester 1986/ 1987, nach einer fünfjährigen Erprobungszeit durch das Institut für Test- und Begabungsforschung (ITB) unter der Leitung des Direktors Dr. Tost, obligatorisch der „Test für medizinische Studiengänge“ (TMS) in den Studiengängen der Human-, Tier- und Zahnmedizin eingeführt. Die Teilnahme an diesem Test war für alle Personen verbindlich, die sich für einen solchen Studienplatz bewerben wollten. Schülerinnen und Schüler des Gymnasiums oder vergleichbarer Einrichtungen konnten den Test bereits im letzten Jahr ihrer Schulausbildung absolvieren. Das Ergebnis des TMS wurde bei der Studienplatzvergabe durch die Zentralstelle für die Vergabe von Studienplätzen (ZVS) berücksichtigt. Im Unterschied zum früheren Zulassungsverfahren konnten die Aussichten auf eine rasche Zulassung zum Studium, die durch ein gutes Abiturzeugnis eröffnet worden wären, durch eine nur mäßige Leistung im TMS zunichte gemacht werden. Umgekehrt konnten allerdings Schwächen im Abiturzeugnis durch eine hohe Leistung im TMS ausgeglichen werden. Letzteres stellte somit eine zusätzliche Chance für Medizininteressenten dar, die nur ein mäßiges Abiturzeugnis hatten. Der TMS sollte Fähigkeiten und Fertigkeiten prüfen, die für eine erfolgreiche Bewältigung der Anforderungen in den Studiengängen Human-, Tier- und Zahnmedizin wichtig sind. Eine zentrale Studienvoraussetzung ist das Verständnis für naturwissenschaftliche und medizinische Problemstellungen, das mit Hilfe des Tests überprüft werden sollte. Zur Beurteilung des Studienerfolges wurde das Abschneiden bei der Ärztlichen Vorprüfung herangezogen.

1996 wurde der TMS aufgrund der „Normalisierung der Zustände" abgeschafft und der Studienantritt mit einem Numerus Clausus nach Schulnoten vereinheitlicht. Während der Zeit, in der der Test verwendet wurde, haben mehr als 300.000 Personen an ihm teilgenommen. In der Schweiz wird seit 1998 der Eignungstest für medizinische Studiengänge (EMS), eine Adaption des TMS, zur Auswahl der Studierenden herangezogen.

In dieser Hausarbeit soll ein Überblick über die Historie des TMS, den Testaufbau, die Durchführung, die Auswertung und die damit verbundene Vorhersehbarkeit des Studienerfolges dargestellt werden.

Historie

Ende der siebziger Jahre wurde nach dem Urteil des Bundesverfassungsgerichtes vom Jahr 1977, das am bisherigen Zulassungsverfahren Kritik geübt hatte, der Einsatz von psychologischen Tests von den in Bund und Ländern Verantwortlichen erwogen. In einer gemeinsamen Initiative von Bund und Ländern wurden eine Reihe von Wissenschaftlern (Psychologen und Erziehungswissenschaftler einerseits, Mediziner andererseits) um Stellungnahmen gebeten. Aus diesem Kreis wissenschaftlicher Berater sind dann Empfehlungen hervorgegangen, nach denen das Institut für Test- und Begabungsforschung (ITB) unter der Leitung des Direktors Dr. Tost mit der Entwicklung eines Hochschulzugangstests für medizinische Studienfächer beauftragt wurde. Nach etwa zwei Jahren der Testentwicklung kam das ITB zu dem Ergebnis, dass das Verfahren im Prinzip einsatzfähig ist. Die Entwicklung des TMS beruht auf einer Reihe von Vorarbeiten zur Beschreibung des Konstrukts „Fähigkeit zur Bewältigung der Anforderungen des Medizinstudiums“. Hierfür wurden die folgenden Arbeiten berücksichtigt:

- Analyse von Erfahrungen bei der Konstruktion eines vergleichbaren Tests in den USA, des „Medical College Admission Test“ (MCAT) (MCAT; Erdmann, Mattson, Hutton & Wallace, 1971; Asscociation of American Medical Colleges, 1976)
- Erstellung und Validierung vorläufiger Anforderungsprofile für die medizinischen Studiengänge (Amelang, 1975; Amelang und Hoppensack, 1976; Hitpass, 1975)
- Empirische Ausbildungsplatzanalysen in den Studiengängen Human, Tier- und Zahnmedizin (Görtzen, 1977; Hoyos & Görtzen, 1976; Intraplan, 1977)
- Einer Befragung von über 100 Lehrenden der Medizin und medizinrelevanter Naturwissenschaften zu dem im Studium erforderlichen kognitiven Fähigkeiten

In den Jahren 1980 bis 1985 wurde in der Bundesrepublik Deutschland der „Test für medizinische Studiengänge“ (TMS) als Ausleseinstrument im Rahmen der Zulassung zu den Studiengängen Human-, Tier- und Zahnmedizin erprobt und in dieser Zeit wurden bereits umfangreiche Informationen zur prognostischen Validität des TMS gewonnen. Sie bezogen sich allerdings auf den Test in der Struktur, die er während dieser Zeit hatte. Damals bestand der Test aus 13 Aufgabengruppen, die Anzahl der Aufgaben pro Untertest war etwas niedriger und zu jedem Testtermin war der TMS in 6 Testformen angelegt (später 8). Nach Analyse der Daten wurden aus den 13 Aufgabengruppen 9 ausgewählt, die sich als prognostisch am ergiebigsten erwiesen hatten. Es durfte damit gerechnet werden, dass der TMS in seiner revidierten Form zumindest zu der gleichen Prognosekraft gelangt wie sein Vorgänger in der Erprobungsphase. Diese Überprüfung wurde in einer großen Längsschnittstudie im Zeitraum von 1986 bis Ende des Jahres 1992 vorgenommen.

Testaufbau des TMS

Der TMS besteht aus neun verschiedenen Untertests, die zu einem Gesamtwert verrechnet werden. Die folgende Tabelle gibt einen Überblick über den Testaufbau, die Aufgabenzahl und die Dauer der einzelnen Untertests:

Tabelle 1: Struktur und Ablauf des TMS; Aufgabenzahl und in Klammern die Zahl der gewerteten Aufgaben pro Unter- und Gesamttest, da zusätzliche (nicht gewertete) Einstreuaufgaben verwendet worden sind, Erläuterungen siehe Text. (Quelle: Trainingsversionzum TMS, 1988)

Bezeichnung der Untertests	**Geprüfte Fähigkeiten**	**Zahl der Aufgaben**	**Bearbeitungszeit (in Minuten)**
Muster zuordnen	Differenzierte visuelle Wahrnehmung	24 (20)	22
Medizinisch-naturwissenschaftliches Grundverständnis	Verständnis für medizinisch-naturwissenschaftliche Problemstellungen	24 (20)	60
Schlauchfiguren	Räumliches Vorstellungsvermögen	24 (20)	15
Quantitative und formale Probleme	Quantitatives Problemlösen in medizinisch-naturwissenschaftlichen Kontexten	24 (20)	60
Konzentriertes und sorgfältiges Arbeiten	Konzentrationsfähigkeit, Aufmerksamkeit	20	8
M I T T A G S P A U S E 60 Minuten			
Lernphase zu den Gedächtnistests:			
Figuren lernen		20	4
Fakten lernen		15	6
Textverständnis	Verständnis und Interpretation medizinischer und naturwissenschaftlicher Texte	24 (18)	60
Reproduktionsphase:			
Figuren lernen	Behalten von figuralem Material	20	5
Fakten lernen	Behalten von verbalem Material	20	7
Diagramme und Tabellen	Interpretation von Diagrammen und Tabellen	24 (20)	60
Gesamttest		**204**	**ca. 5 Std. zzgl. Pause**

Jedes Jahr wurden neue Aufgaben für die Untertests entwickelt und in mehreren Schritten überarbeitet. An der Aufgabenentwicklung nahmen zahlreiche Lehrbeauftragte und Experten teil. Die Aufgaben müssen sehr hohe Qualitätsstandards erfüllen, u.a.

- müssen sie jedes Jahr die Studieneignung gleich zuverlässig messen
- muss das Schwierigkeitsspektrum aller Aufgaben annähernd vergleichbar sein
- darf kein spezielles Fachwissen vorausgesetzt werden, um die Trainierbarkeit des Tests gering zu halten
- muss eine eindeutige richtige Lösung existieren

Die Erprobung neuer Aufgaben für sechs der neun Untertests (siehe Tabelle 1) erfolgte im Rahmen sogenannter „Einstreuaufgaben“. Nur bei ausreichender Bewährung wurden solche Aufgaben dann in nachfolgenden Testversionen für die Werteberechnung verwendet. Vier neue Aufgaben pro Untertest wurden in jeder Testform probeweise mitbearbeitet– ihr Ergebnis wurde jedoch nicht gezählt. Da 8 verschiedene Testformen bei jedem Durchführungstermin zusammengestellt wurden, konnten jeweils 32 neue Aufgaben pro Untertest an ausreichend großer Stichproben untersucht werden. Maximal 20 davon wurden pro Jahr gebraucht– dieser Überschuss war nach Aussagen der Entwickler auch notwendig, da nicht alle Aufgaben die Kriterien zufriedenstellend erfüllten. Die Einstreuaufgaben wurden nicht besonders gekennzeichnet- jede Aufgabe des Tests konnte eine solche sein.

Beispielaufgaben für die Untertests

Nachfolgend wird pro Untertest eine Beispielaufgabe dargestellt. Diese Beispielaufgaben sind aus der Originalversion des Tests für medizinische Studiengänge vom Institut für Test- und Begabungsforschung von 1990 entnommen. So kann das Prinzip der Aufgabenstruktur verdeutlicht werden– die Aufgaben unterscheiden sich innerhalb jedes Untertests bezüglich des Schwierigkeitsgrades und der Anforderung beträchtlich. Auch die Vielfalt der Aufgaben ist sehr groß und nur sehr schwer darstellbar. Einen besseren Überblick geben die kompletten Originalversionen, die auch im Seminar vorgestellt werden.

Bei der Beurteilung der Aufgaben wird die Nähe zu Studienanforderungen deutlich: an der Entwicklung haben neben Gymnasiallehrern, Medizinern und Psychologen auch Lehrbeauftragte des Grundstudiums Medizin mitgearbeitet. Die Struktur der Untertests ist auf detaillierte Anforderungsanalysen eines Medizinstudiums zurückzuführen, die im Rahmen der Testentwicklung durchgeführt worden sind (Trost, 1997).

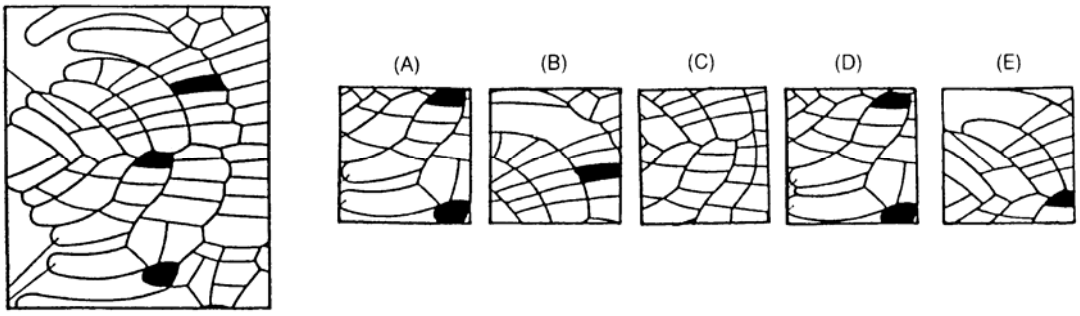
Untertest: Muster zuordnen

In diesem Untertest wird die Fähigkeit geprüft, Ausschnitte in einem komplexen Bild wiederzuerkennen. Dazu werden pro Aufgabe ein Muster und je fünf Musterausschnitte (A) bis (E) vorgegeben. Die Testteilnehmerin oder der Testteilnehmer soll herausfinden,

welcher dieser fünf Musterausschnitte an irgendeiner beliebigen Stelle deckungsgleich und vollständig auf das Muster gelegt werden kann.

Ein Beispiel dazu:

Muster Musterausschnitte



In den meisten Aufgaben dieser Art heben sich die vier nicht deckungsgleichen Musterausschnitte dadurch vom Muster ab, dass Details entweder hinzugefügt oder weggelassen sind. Zugleich stellt dieser Untertest Anforderungen an die Schnelligkeit der Bearbeitung.

In durchschnittlich 55 Sekunden je Aufgabe muss die Testperson die richtige Lösung herausgefunden haben, dass beispielsweise in der obigen Aufgabe nur der Musterausschnitt (A) deckungsgleich mit einem Teil des Musters ist, und zwar in dessen unterem Bereich, etwa in der Mitte.

Untertest: Medizinisch-naturwissenschaftliches Grundverständnis

Hier wird das Verständnis für Fragen der Medizin und der Naturwissenschaften geprüft. Der Text könnte so in einem Lehrbuch stehen. Wichtig für das Verständnis dieser Textpassage ist, ob daraus bestimmte logische Schlüsse gezogen werden können. Alle Fakten, die für die Beantwortung der Aufgabe notwendig sind, stehen im Text– spezielles medizinisches Vorwissen ist nicht erforderlich. Dieses wichtige Prinzip findet sich bei allen Untertests und ist verantwortlich für die gewünschte geringe Trainierbarkeit der Aufgabenlösung.

Beispielaufgabe:

Im Kindesalter kann das Zentrum für Sprache, Spracherwerb und Sprachverständnis noch in der linken oder in der rechten Hälfte (Hemisphäre) des Gehirns in einem umschriebenen Hirnrindengebiet (sog. Sprachregion) angelegt werden. Spätestens im zwölften Lebensjahr sind die sprachlichen Fähigkeiten jedoch fest in einer der beiden Hemisphären verankert, und zwar bei den Rechtshändern in der Regel links, bei den Linkshändern in der Mehrzahl ebenfalls links, zum Teil aber auch rechts; die korrespondierende Region der Gegenseite hat

zu diesem Zeitpunkt bereits andere Funktionen fest übernommen. Welche der nachfolgenden Aussagen lässt bzw. lassen sich aus diesen Informationen ableiten?

Bei irreversiblen Hirnrindenverletzungen im Bereich der sogenannten Sprachregion der linken Hemisphäre ...

I. kommt es bei erwachsenen Linkshändern in der Regel zu keinen wesentlichen Sprachstörungen.

II. kommt es bei einem Vorschulkind in der Regel zu einer bleibenden Unfähigkeit, die Muttersprache wieder zu erlernen.

III. ist bei zwanzigjährigen Rechtshändern die Fähigkeit, eine Sprache zu erlernen, in der Regel verloren gegangen.

(A) Nur Ausfall I ist zu erwarten.

(B) Nur Ausfall II ist zu erwarten.

(C) Nur Ausfall III ist zu erwarten.

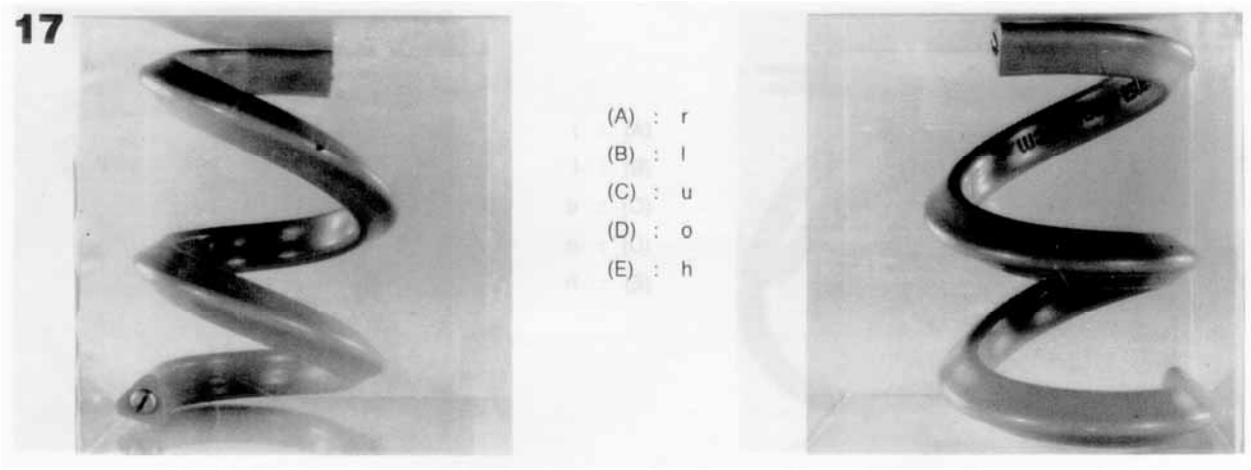
(D) Nur die Ausfälle I und III sind zu erwarten.

(E) Nur die Ausfälle II und III sind zu erwarten.

Untertest: Schlauchfiguren

Diese Aufgaben sollen das räumliche Vorstellungsvermögen prüfen– eine Funktion, die beispielsweise für das Verständnis von Röntgenbildern wichtig ist. Während des Studiums werden zahlreiche eigentlich dreidimensional zu betrachtende Strukturen und Vorgänge in zweidimensionalen Abbildungen vermittelt.

Jede Aufgabe besteht aus zwei Abbildungen eines durchsichtigen Würfels, in dem sich ein, zwei oder drei Kabel befinden. Die erste Abbildung (links) zeigt stets die Vorderansicht des Würfels; auf dem rechten Bild daneben, wo derselbe Würfel noch einmal abgebildet ist, soll die Testteilnehmerin oder der Testteilnehmer herausfinden, ob die Abbildung die Ansicht von rechts (r), links (l), unten (u), oben (o) oder von hinten (h) zeigt.



Hier sehen Sie den Würfel von vorne! Hier sehen Sie den Würfel von? (hinten!)

Untertest: Quantitative und formale Probleme

Mit Hilfe dieses Untertests wird die Fähigkeit überprüft, im Rahmen medizinischer und naturwissenschaftlicher Fragestellungen mit Zahlen, Größen, Einheiten und Formeln richtig umzugehen. Diese Anforderung ist für mehrere Fächer des Grundlagenstudiums der Medizin bedeutsam. Bei solchen Fragen werden die Kenntnisse der Mittelstufenmathematik vorausgesetzt.

Beispielaufgabe:

Aus einer 10%igen NaCl-Stammlösung soll durch Verdünnung eine 0,5%ige NaCl-Lösung hergestellt werden. Auf welches Volumen muss mit Wasser aufgefüllt werden, wenn man 1 ml der Stammlösung nimmt?

(A) 10 ml

(B) 15 ml

(C) 20 ml

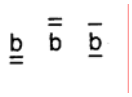
(D) 40 ml

(E) 200 ml

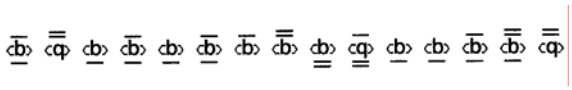
Untertest: Konzentriertes und sorgfältiges Arbeiten

Bei diesem Untertest soll die Fähigkeit, rasch, sorgfältig und konzentriert zu arbeiten, gemessen werden. Dabei sollen möglichst alle b, die mit zwei Querstreichen versehen sind, die entweder beide unten, beide oben oder je einer unten und oben angebracht sind, markiert werden. Die Lösungsmenge ist ebenso wichtig wie die Fehlerfreiheit der Bearbeitung. Dieser Test ist trainierbar. Es wird vorab in der Test-Information darauf

hingewiesen, diesen Untertest vor der Testabnahme mehrfach zu üben, um ein gutes Ergebnis zu erzielen.



Diese Buchstaben b mit zwei Querstrichen sind eingestreut unter b mit einem, drei oder vier Querstrichen sowie unter q mit einem oder mehreren Querstrichen. Im folgenden Beispiel wären also das 1., 4., 6., 8., 9. und 13. Zeichen zu markieren.



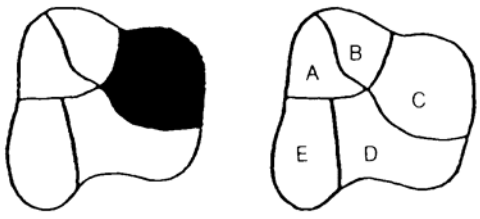
Untertest: Figuren lernen

Für beide nachfolgenden Gedächtnistests wird nach der Mittagspause das Material zum Einprägen ausgeteilt. Vor der Abfrage des Gelernten wird der Untertest „Textverständnis“ bearbeitet und damit liegt die Zeit bis hin zur Reproduktion bei über einer Stunde. Gedächtnisleistungen werden als wichtige Voraussetzungen für Studienerfolg bewertet.

Der Untertest „Figuren lernen“ prüft, wie gut man sich Einzelheiten von Gegenständen einprägen und merken kann.

Ein Beispiel:

Gezeigte Figur zum Einprägen Gezeigte Figur beim Abfragen



Die Testperson hat vier Minuten Zeit, um sich 20 solcher Figuren einschließlich der Lage der schwarzen Flächen einzuprägen. Nach ca. einer Stunde müssen die Testpersonen angeben können, welcher Teil der Abbildung markiert war, um dies auf dem Antwortbogen einzutragen. Die Lösung ist Fläche C.

Untertest: Fakten lernen

Analog dem Prinzip beim "Figuren lernen" sollen hier Fakten eingeprägt und behalten werden, die ebenfalls nach der gleichen Zwischenzeit abgefragt werden. Dabei werden 15

Patienten vorgestellt, von denen jeweils der Name, die Altersgruppe, Beruf und Geschlecht, ein weiteres Beschreibungsmerkmal (z.B. Familienstand) sowie die Diagnose erfahren wird. Ein Beispiel für eine derartige Fallbeschreibung ist:

Lemke, 30 Jahre, Dachdecker, ledig, Schädelbasisbruch

Eine Frage zum obigen Beispiel könnte z.B. lauten:

Der Patient mit dem Schädelbruch ist von Beruf ...

(A) Installateur

(B) Lehrer

(C) Dachdecker

(D) Handelsvertreter

(E) Physiker

Untertest: Textverständnis

Dieser Untertest soll die Fähigkeit überprüfen, umfangreiches und komplexes Textmaterial aufzunehmen und zu verarbeiten. Die Texte sind inhaltlich und grammatikalisch anspruchsvoll– sie können unter Nutzung von Notizen und Unterstreichungen erarbeitet werden. Die Überprüfung des Textverständnisses erfolgt über die Auswahl einer richtigen oder falschen Aussage aus fünf vorgegebenen Antworten.

Beispielaufgabe:

Zu den Aufgaben der Schilddrüse gehören Bildung, Speicherung und Freisetzung der jodhaltigen Hormone Trijodthyronin (T_3) und Thyroxin (T_4). In der Schilddrüse befinden sich zahlreiche Hohlräume, Follikel genannt, deren Wände von einer Schicht sogenannter Epithelzellen gebildet werden. Diese Follikel sind mit einer Substanz gefüllt, in der die Hormone T_3 und T_4 als inaktive Speicherformen enthalten sind. Beim Menschen ist in den Follikeln so viel T_3 und T_4 gespeichert, dass der Organismus damit für etwa 10 Monate versorgt werden kann.

Das für die Hormonbildung erforderliche Jod entstammt der Nahrung und wird von den Epithelzellen als Jodid aus dem Blut aufgenommen. Die Jodidaufnahme erfolgt an der äusseren Zellmembran der Epithelzellen durch eine sogenannte Jodpumpe. Diese wird durch ein Hormon aus der Hirnanhangsdrüse, das TSH, stimuliert und kann

pharmakologisch durch die Gabe von Perchlorat gehemmt werden. Ferner gibt es erbliche Schilddrüsenerkrankungen, bei deren Vorliegen die Jodpumpe nicht funktioniert.

Bei Gesunden wird das in die Epithelzellen aufgenommene Jodid im nächsten Schritt unter dem Einfluss eines Enzyms in freies Jod umgewandelt und in die Follikel abgegeben. Die Aktivität dieses Enzyms kann ebenfalls pharmakologisch gehemmt werden.

Die letzten Schritte der Hormonbildung finden in den Follikeln, also ausserhalb der einzelnen Epithelzellen, statt. In dort vorhandene sogenannte Tyrosin-Reste (des Thyreoglobulins) wird zunächst ein Jodatom eingebaut. So entstehen Monojodtyrosin-Reste (MIT), von denen ein Teil durch die Bindung je eines weiteren Jodatoms in Dijodtyrosin-Reste (DIT) umgewandelt wird. Durch die Verknüpfung von je zwei DIT-Resten entsteht schliesslich T_4, während aus der Verbindung je eines MIT-Restes mit einem DIT-Rest T_3 hervorgeht. T_3 und T_4 werden dann in den Follikeln gespeichert und bei Bedarf über die Epithelzellen ins Blut freigesetzt.

Diese Freisetzung von T_3 und T_4 ins Blut (Sekretion) wird über die Hirnanhangsdrüse und den Hypothalamus, einen Teil des Zwischenhirns, gesteuert: Das erwähnte Hormon TSH stimuliert ausser der Bildung auch die Sekretion von T_3 und T_4; es ist hinsichtlich seiner eigenen Sekretionsrate jedoch abhängig von der Stimulation durch das hypothalamische Hormon TRH. Die TRH-Sekretion wiederum wird z.B. durch Kälte stimuliert, während Wärme hemmend wirken kann. Neben diesen übergeordneten Steuerungsmechanismen existiert noch ein sogenannter Rückkoppelungsmechanismus: Eine hohe Konzentration von T_3 und T_4 im Blut hemmt die TSH- und die TRH-Sekretion, eine niedrige Konzentration stimuliert sie. Bei den an der Steuerung der Schilddrüsenhormon-Sekretion beteiligten Arealen von Hirnanhangsdrüse und Hypothalamus können krankheitsbedingte Störungen auftreten, die zu einer Über- oder Unterfunktion der Schilddrüse führen.

Eine der Hauptwirkungen von T_3 und T_4 ist die Beeinflussung des Energieumsatzes durch eine Steigerung des Sauerstoffverbrauchs in stoffwechselaktiven Organen. Entsprechend senkt eine zu niedrige Konzentration der beiden Hormone im Blut (Hypothyreose) den Energieumsatz bzw. die Stoffwechselaktivität unter den normalen Wert, während bei einer zu hohen Konzentration (Hyperthyreose) die Stoffwechselaktivität gesteigert wird. Die Hormone T_3 und T_4 können ebenso wie TSH und TRH für diagnostische und therapeutische Zwecke synthetisch hergestellt werden.

Auf einen solchen Text folgen Fragen, die sich ausschließlich auf im Text vorhandene Inhalte beziehen; eine Frage mit niedrigem Schwierigkeitsgrad ist zum Beispiel so formuliert:

Welcher der folgenden Vorgänge gehört nicht zu den im Text beschriebenen Schritten, die zur Bildung von T_3 führen?

(A) Transport von Jod aus den Epithelzellen in die Follikel

(B) Umwandlung von Jod in Jodid in den Follikeln

(C) Transport von Jodid aus dem Blut in die Epithelzellen

(D) Verknüpfung von MIT- und DIT-Resten in den Follikeln

(E) Verknüpfung von Jod und Tyrosin-Resten in den Follikeln

Untertest: Diagramme und Tabellen

Mit dieser Aufgabengruppe wird die Fähigkeit geprüft, Diagramme und Tabellen richtig zu analysieren und zu interpretieren. In dieser Form werden während des Studiums zahlreiche Zusammenhänge vermittelt. Wie bei den Untertests „Medizinisch-naturwissenschaftliches Grundverständnis" und „Textverständnis" sind auch hier zur Lösung dieser Aufgabe keine speziellen naturwissenschaftlichen, medizinischen oder statistischen Kenntnisse erforderlich.

Eine Aufgabe dazu:

Die folgende Tabelle beschreibt die Zusammensetzung und den Energiegehalt von vier verschiedenen Milcharten. Unter Energiegehalt der Milch verstehen wir dabei die Energiemenge, gemessen in Kilojoule (kJ), welche 100 Gramm (g) Milch dem Organismus ihres Konsumenten liefern können.

Milchart	**Eiweiss**	**Fett**	**Milchzucker**	**Salze**	**Energiegehalt**
menschliche Muttermilch	1,2 g	4,0 g	7,0 g	0,25 g	294 kJ
Vollmilch	3,5 g	3,5 g	4,5 g	0,75 g	273 kJ
Magermilch	3,3 g	0,5 g	4,5 g	0,75 g	160 kJ
Buttermilch	3,0 g	0,5 g	3,0 g	0,55 g	110 kJ

Welche Aussage lässt sich aus den gegebenen Informationen nicht ableiten?

(A) Menschliche Muttermilch enthält mehr als doppelt soviel Milchzucker wie Buttermilch.

(B) Vollmilch enthält im Vergleich zur menschlichen Muttermilch etwa die dreifache Menge an Salzen und Eiweiß.

(C) Zur Aufnahme der gleichen Energiemenge muss ein Säugling fast dreimal soviel Buttermilch wie Muttermilch trinken.

(D) Der Unterschied zwischen Magermilch und Vollmilch ist bei der Mehrzahl der aufgeführten Merkmale geringer als der Unterschied zwischen Magermilch und Buttermilch.

(E) Der Eiweißgehalt der Milch ist für den Energiegehalt von entscheidender Bedeutung.

Berechnung der Werte

Alle Untertests außer dem „Konzentrierten und sorgfältigen Arbeiten" liefern eine Summe („Punkte") richtig gelöster Aufgaben zwischen 0 und 20 bzw. 18 (bei „Textverständnis“). Summiert werden die gewerteten Aufgaben, nicht die Einstreuaufgaben.

Beim Test „Konzentriertes und sorgfältiges Arbeiten“ müssen insgesamt 1200 Zeichen der Reihe nach bearbeitet werden– 600 davon sind anzustreichen. Es können in der zur Verfügung stehenden Zeit in der Regel nicht alle Zeichen bearbeitet werden. Die Position des letzten angestrichenen Zeichens bestimmt, wie viele Zeichen als bearbeitet gewertet werden. Alle übersehenen und fälschlich angestrichenen Zeichen vor diesem letzten bearbeiteten Zeichen zählen als Fehler und diese werden von der Menge der insgesamt angestrichenen Zeichen abgezogen. Die verbleibende Menge sind die „Richtigen", die dann in eine Skala zwischen 0 und 20 transformiert werden, um mit den anderen Tests gleichgewichtig zum Punktwert addiert zu werden. 600 „Richtige“ wären das Maximum und entsprechen 20 Punkten. Alle Punkte der Untertests werden zu einer Summe addiert (Punktwert, vgl. Abbildung 1). Dieser Wert hat den Nachteil, dass er nicht zwischen Tests verschiedener Jahrgänge vergleichbar ist. Es findet eine Standardisierung auf den Mittelwert und die Standardabweichung der jeweiligen Testform statt. Dieser Testwert liegt zwischen 70 und 130 (der Mittelwert ist 100) und kann in einen Prozentrangwert umgerechnet werden.

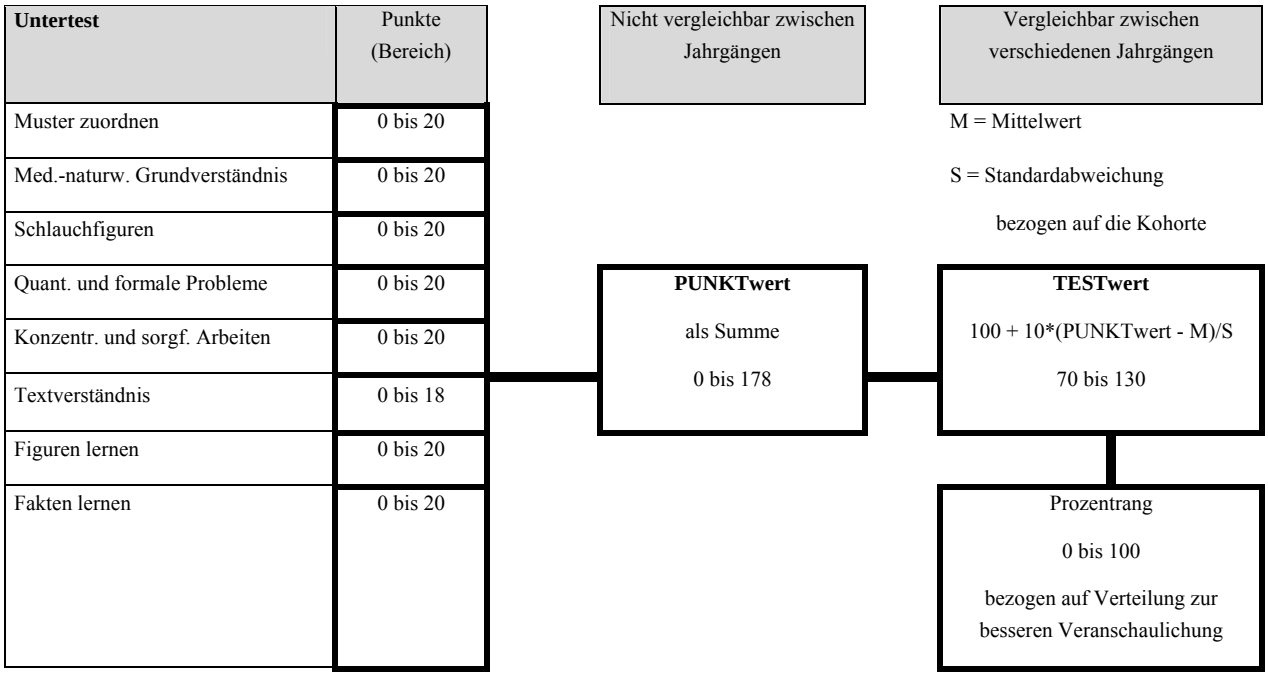


Abbildung 1: Punktwerte der einzelnen Untertests und ihre Zusammenführung über den Punktwert zum Testwert.

Vorhersage des Studienerfolges durch den TMS

Der Test wurde in Deutschland über zehn Jahre angewendet. In dieser Zeit fanden mehrere Evaluationen statt, um seine Gütekriterien zu überprüfen (vgl. Bericht von Trost u.a., 1998). Die Vorhersagefähigkeit von Studienerfolg (gemessen an der Ärztlichen Vorprüfung) ist dabei das zentrale Kriterium. Mit Hilfe des Tests sollten diejenigen bevorzugt eine Chance erhalten, die mit größerer Wahrscheinlichkeit das Studium auch erfolgreich beenden. Die folgenden Ergebnisse sind dem 18. Arbeitsbericht des Instituts für Test- und Begabungsforschung (Trost, 1994) entnommen. Zu beachten ist, dass es in Deutschland fünf Zugangsmöglichkeiten, prozentual verteilt, zum Studium gab:

- 45 % durch eine kombinierte Abitur-Test-Quote (Kriterium ist die Kombination von Abiturdurchschnittsnote und Testergebnis im Verhältnis 55 zu 45)
- 10% durch nur die Test-Quote (Kriterium ist allein das Testergebnis)
- 20% durch Wartezeit (Kriterium ist Anzahl der Semester, in denen ein Kandidat sich für den betreffenden Studiengang beworben hat)
- 15% aufgrund von Auswahlgesprächen (Kriterium ist das Ergebnis in einem Auswahlgespräch, das der Kandidat mit zwei Hochschullehrern führt)
- 10% als „Vorab-Quote" (z.B. bereits früher zugesagte Plätze, bestimmte Gruppe von Ausländern, Zweitstudienplätze, Härtefälle)

Die Berücksichtigung der Quoten erfolgt sequentiell: In diesem „Kaskaden- Modell“ erfolgt die Auswahl nach der Abitur-Test-Quote zuerst, dann die nach der Testquote. Personen mit guten Abitur- und Testleistungen werden also bereits in der ersten Quote berücksichtigt. Die Test- Quote beinhaltet dann Personen, die bei Kombination Abitur und Test nicht zugelassen werden können, aber eine gute Testleistung erreicht haben. Die Auswahlgesprächs-Quote trifft dann nur für Personen zu, die nicht mittels einer der erstgenannten zugelassen worden sind.

Zur Beurteilung des Studienerfolges wurden folgende Indikatoren gewählt:

- Kürze der Studiendauer bis zum erstmaligen Antreten zur Ärztlichen Vorprüfung (nach Studienordnung frühestens nach 4 Fachsemestern)
- Bestehen der Ärztlichen Vorprüfung beim ersten Versuch
- erreichte Punktzahl bzw. Noten in der Ärztlichen Vorprüfung insgesamt, im schriftlichen/ mündlichen Prüfungsteil sowie in den einzelnen Prüfungsfächern

In die Beurteilung der prognostischen Validität des TMS wurden alle Personen einbezogen,

- die an einem der ersten drei Testterminen (1986, 1987) teilgenommen hatten
- die in der Folgezeit, über welche Quote auch immer, zum Studiengang Medizin zugelassen worden waren und
- die bis zum Herbst 1992 zur Ärztlichen Vorprüfung angetreten waren (Nauels & Klieme, 1994a)

Trost beschreibt, dass aus zahlreichen Untersuchungen bekannt ist, dass Studierende, die nach vergleichsweise kurzer Studiendauer zur Prüfung antreten, in der Regel bessere Prüfungsleistungen erzielen. Aus diesen auch ökonomischen Gesichtspunkten wurde für die prognostische Validität als erstes Kriterium die Semesterzahl bis zum Antreten der Ärztlichen Vorprüfung gewählt. Das Ergebnis zeigt, dass nach vier Fachsemestern 85% der über die Abitur-Test-Quote, 72% der über die Test-Quote, 61% der über die Wartezeit-Quote und 65% der über die Auswahlgesprächs-Quote Zugelassenen zur Ärztlichen Vorprüfung antraten. 3,5% der Abitur-Test-Quote und 15% der über Wartezeit-Quote Zugelassenen stellte sich erst nach sechs oder mehr Fachsemestern.

Ein noch deutlicherer Unterschied zeigt sich, wenn man als Indikator für den Studienerfolg das Bestehen der Ärztlichen Vorprüfung im ersten Versuch heranzieht. Hier kann man folgende prozentuale erfolgreiche Teilnahme feststellen:

- 92% der über die Abitur-Test-Quote
- 81% der über die Test-Quote
- 64% der über die Wartezeit-Quote
- 67% der über die Auswahlgesprächs-Quote Zugelassenen

Verifiziert man das Bestehen noch danach, dass man vorgibt, dass nach der Mindest-Studiendauer bestanden worden ist, ergibt sich innerhalb der Gruppen folgendes Ergebnis:

- 80% der über die Abitur-Test-Quote
- 62% der über die Test-Quote
- 45% der über die Wartezeit-Quote
- 49% der über die Auswahlgesprächs-Quote Zugelassenen

In der schriftlichen Prüfung (Abbildung 2) werden von den Personen mit Zulassung nach der Abitur-Test- und der Testquote die besten Leistungen erzielt. Bemerkenswert ist, dass die Personen der Auswahlgesprächs-Quote hier die schlechtesten Leistungen erreichen. Auch in der mündlichen Prüfung zeigt sich dieser Trend.

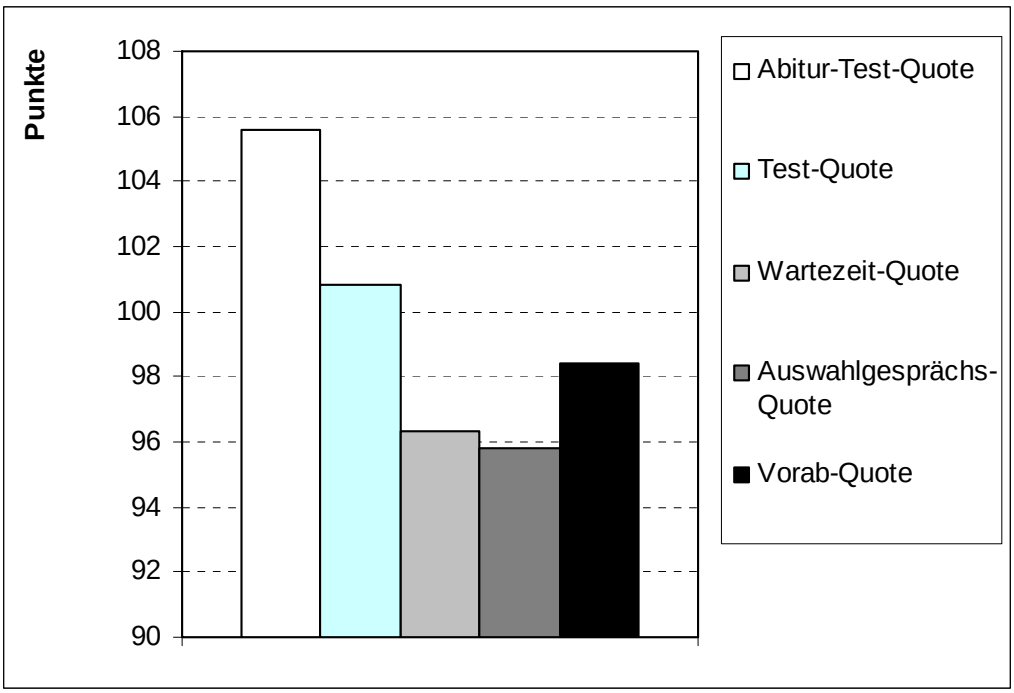


Abbildung 2: Durchschnittliche Punkte für schriftliche Prüfungsleistungen während des Studiums nach der Zulassung (aus Trost, 1994, S. 144).

Betrachtet man abschließend noch einmal die Erfolgsrate der Studierenden in der Ärztlichen Vorprüfung beim ersten Anlauf lag diese bei einem Zulassungsverfahren über Los bei 69%. Innerhalb der untersuchten Gruppe lag indessen die Erfolgsrate derjenigen Teilgruppen, die durch die Kombination von Abiturnote und Testergebnis zugelassen wurden, bei 96%. Die Veränderung des Auswahlverfahrens weg vom Zufallverfahren bedeutete eine Verbesserung der Erfolgsrate um 27 Prozentpunkte.

In den Untersuchungen von Trost wurde des Weiteren untersucht in wieweit nicht schon die Abiturnote Aufschluss über die Studienleistung geben kann. Die Ergebnisse dieser Untersuchungen zeigten, dass der TMS die Abiturdurchschnittsnote an Prognosekraft

hinsichtlich des schriftlichen Gesamtergebnisses übersteigt. Außerdem wurde festgestellt, dass der Prognosewert in der Kombination von Testergebnis und Abiturdurchschnittsnote gemeinsam deutlich höher war als der Prognosewert eines jeden einzelnen Auswahlkriteriums allein. Vergleicht man das sofortige Bestehen der Ärztlichen Vorprüfung zwischen „Abitur-Besseren“ und „Abitur-Schwächeren“ lag die Bestehensrate 92% zu 70%. Bei den „Test-Besseren“ zu den „Test-Schwächeren“ zeigte sich ein ähnliches Ergebnis bei 91% zu 70%. In der Kombination von Abiturnote und Testwert lagen die Erfolgsraten bei 93% zu 69%.

Die Korrelationen in Tabelle 2 verdeutlichen, dass vor allem die schriftliche Prüfungsnote durch den Test sehr gut vorhergesagt werden kann. Man kann an den unterschiedlichen Korrelationen erkennen, dass bestimmte Untertests des TMS eine klare und auch über die „Test-Kohorten“ hinweg recht stabile Prognose über den Studienerfolg geben konnten.

Tabelle 2: Korrelationen zwischen den Prädiktorvariablen (TMS-Ergebnisse) und den Ergebnissen der ärztlichen Vorprüfungen; aus Trost (1994).

	Korrelationen mit den Ergebnissen der Ärztlichen Vorprüfungen		
Prädiktorvariable	**schriftlicher Teil**	**mündlicher Teil**	**Gesamtnote**
TMS-Gesamtwert	**0.50**	**0.27**	**0.45**
Muster zuordnen	0.17	0.10	0.14
Med.-naturw. Grundverständnis	0.41	0.21	0.36
Schlauchfiguren	0.28	0.14	0.25
Quantitative und formale Probleme	0.45	0.21	0.39
Konzentriertes und sorgf. Arbeiten	0.24	0.16	0.22
Textverständnis	0.41	0.21	0.35
Figuren lernen	0.22	0.14	0.21
Fakten lernen	0.17	0.11	0.16
Diagramme und Tabellen	0.41	0.21	0.37

Zusammenfassung

Zusammenfassend kann man feststellen, dass der „Test für medizinische Studiengänge“ (TMS) die Erwartungen, die an ihn gestellt wurden, erfüllt hat. Hauptkriterium war definiert als die Vorhersagefähigkeit des Studienerfolges im Studienfach der Human-, Tier- und Zahnmedizin. Der Studienerfolg wurde an der Dauer bis zum Erreichen der Ärztlichen Vorprüfung und deren Bestehen bzw. der Leistung beurteilt. Als entscheidendes Kriterium wird die Kombination von Abiturdurchschnittsnote und Testergebnis des TMS benannt. Diese Kombination erhöht die Erfolgsprognose erheblich gegenüber nur einem Merkmal.

Es konnte festgestellt werden, dass Studierende, die über die Abitur-Test-Quote zu einem Medizinstudienplatz zugelassen wurden, im Mittel nicht nur schneller sondern auch im ersten Anlauf die Ärztliche Vorprüfung erfolgreich absolvierten als die über andere Quoten zugelassene Studierenden. Des Weiteren konnte in diesem Zusammenhang festgestellt werden, dass in der schriftlichen Prüfung von den Personen mit Zulassung nach der Abitur-Test- und der Test-Quote die besten Leistungen erzielt wurden. Die Personen der Auswahlgesprächs-Quote erreichten die schlechtesten Leistungen. Auch in der mündlichen Prüfung zeigt sich dieser Trend. Aus ökonomischen Gesichtspunkten kann festgestellt werden, dass der Einsatz des TMS die Prognose bezüglich der Studiendauer sowie des Studienerfolges beeinflusst und damit positive Auswirkungen hatte.

Trotz des scheinbar großen Nutzens lässt sich jedoch die Effizienz des TMS unter dem Gesichtspunkt in Frage stellen, dass bezüglich der bisherigen Auswertungen (hauptsächlich in den Arbeitsberichten von Trost) in bundesdeutschen Zulassungsverfahren die Bewerber, die zu Auswahlgesprächen eingeladen wurden, im Hinblick auf ihre Abitur-Test- und Test-Quote bereits eine "Negativ-Auslese" darstellten. Denn die diesbezüglich stärkeren Bewerber waren bereits in den leistungsbezogenen Quoten zum Zuge gekommen. Im Hinblick auf die generelle Fragestellung nach Studienerfolg in medizinischen Studiengängen wäre es dagegen höchst interessant gewesen, zu erfahren, welche studien- und berufsbezogenen Fähigkeiten- und zwar solche, die nicht in der Ärztlichen Vorprüfung erfasst werden, gerade diese Studierenden aufgezeigt haben.

Literatur

Institut für Test- und Begabungsforschung (Hrsg.) (1988). *Test für medizinische Studiengänge* (Aktualisierte Originalversion 2). Herausgegeben im Auftrag der Kultusminister der Länder der BRD. 2. Auflage. Göttingen: Hogrefe

Institut für Test- und Begabungsforschung (Hrsg.) (1990). *Der neue Test für medizinische Studiengänge* (Originalversion des Tests für medizinische Studiengänge im besonderen Auswahlverfahren). Herausgegeben im Auftrag der Kultusminister der Länder der BRD. 3. Auflage. Göttingen: Hogrefe.

Hayit, E. (1988). *Trainingsversion 1 und 2 zum Mediziner-Test (TMS).* Köln: Hayit Verlag.

Herrmann, T. (1997). *Hochschulentwicklung- Aufgaben und Chancen (Hrsg.).* Heidelberg: Roland Asanger Verlag.

Trost, G. (Hrsg.) (1994). *Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (18. Arbeitsbericht).* Bonn: ITB.

Trost, G. (Hrsg.) (1995). *Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (19. Arbeitsbericht).* Bonn: ITB.

Trost, G. (Hrsg.) (1996). *Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (20. Arbeitsbericht).* Bonn: ITB.

Trost, G. (Hrsg.) (1997). *Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (21. Arbeitsbericht).* Bonn: ITB.

Trost, G., Blum, F., Fay, E., Klieme, E., Maichle, U., Meyer, M. & Nauels, H.-U. (1998). *Evaluation des Tests für Medizinische Studiengänge (TMS): Synopse der Ergebnisse.* Bonn: ITB.

Graduate Record Examination
- General Test -

Sarah Steffens

Einleitung

Die Studierendenauswahl soll zukünftig für zulassungsbeschränkte Fächer sehr viel stärker durch die Hochschulen selbst erfolgen. Da dies in anderen Ländern schon lange so gehandhabt wird, ist es für die Entwicklung eines Auswahlverfahrens von Interesse, sich an bestehenden Tests zu orientieren.

Ein Beispiel für einen Test, der bei der Auswahl in den USA eingesetzt wird, soll im Folgenden behandelt werden. Der Graduate Record Examination (GRE) ist eine Gruppe von standardisierten Tests, die bereits seit mehr als 60 Jahren von vielen Hochschulen und Berufsschulen in den USA von ihren Bewerbern verlangt wird. Dieser zusammengesetzte Test wurde vom *Educational Testing Service (ETS)* entwickelt und wird auch in seinen Zentralen angeboten. Es gibt den GRE als *General Test* und auch für manche Studienfächer als *Subject Test*, der gewöhnlich nach dem Bachelor-Abschluss erfolgen kann. Auf ersteren, den allgemeinen Test, wird hier genauer eingegangen.

Er besteht aus drei verschiedenen Testteilen, welche die Fähigkeit, die Personen in verbalen und quantitativen Aufgaben, sowie in Aufgaben, die analytisches Schreiben verlangen, messen. Dies kann entweder im Papier-Bleistift-Verfahren oder computerbasiert, dann durch adaptives Testen, erfolgen.

Im Folgenden wird auf die Ziele die mit der Testung verfolgt werden, die Struktur des GRE mit seinen verschiedenen Testteilen und den dazugehörigen Aufgabentypen, die Auswertung des Tests, dessen Objektivität, Reliabilität und Studien zur Validität eingegangen. Zum Schluss erfolgen eine Kritik an dem Verfahren sowie eine kurze Zusammenfassung.

Ziele

Der GRE soll Hochschulzulassungs-Kommitees und Stipendiensponsoren helfen, die Qualifikation der Studienbewerber zu erfassen und mit Hilfe dessen, zu entscheiden, ob sie zum Studium zugelassen werden bzw. ein Stipendium erhalten. Die Bewerber können hinsichtlich ihrer Punktzahl sehr gut verglichen und ausgewählt werden. Das Gewicht, das dem GRE-Ergebnis bei der Studierendenauswahl zugesprochen wird, hängt von der jeweiligen Hochschule ab.

Adaptives Testen

Der GRE wird in den meisten Ländern, außer in Amerika, als Papier-Bleistift-Test angeboten. In den USA kann man ihn mittlerweile aber nur noch im CAT-Format (Computer Adaptive Testing) absolvieren. Der Test findet also am Computer statt und passt sich in seiner Schwierigkeit der Fähigkeit des Kandidaten an. In jedem Fall beginnt er mit einer mittelschweren Aufgabe. Beantwortet der Testteilnehmer diese Frage richtig, wird seine Punktzahl erhöht und die nächste Frage wird vom Computer aus einem Itempool ausgewählt. Diese ist dann etwas schwieriger als die erste. Wird die erste Frage falsch beantwortet, so wird die Punktzahl reduziert und das nächste Item ist etwas leichter. Dieser Prozess wiederholt sich, bis der Teilnehmer die maximale Aufgabenzahl bearbeitet hat oder das Zeitlimit für die betreffende Sektion erreicht ist. Jeder Kandidat erhält also eine individuelle Kombination von Fragen. Dadurch wird gewährleistet, dass jeder nur Aufgaben lösen muss, die genau seinem Fähigkeitsniveau entsprechen und somit das höchste Maß an Information über die Teilnehmer bereitstellen (Moosbrugger, 1997). Während beim Papier-Bleistift-Test jede Aufgabe gleich bewertet wird, erhalten die Kandidaten beim CAT mehr Punkte, wenn sie ein schwieriges Item gelöst haben, als wenn es sich um ein leichtes handelte.

Struktur

Die Papier-Bleistift-Form des GRE ist in 4 Sektionen unterteilt und dauert insgesamt 3,75 Stunden. Jeder Testteilnehmer muss einen quantitativen Testteil *(Quantitative Section)*, einen verbalen Testteil *(Verbal Section)*, einen Testteil mit analytischem Schreiben *(Analytical Writing Section)* und einen unidentifizierbaren Vortestteil *(Pretest Section)* bearbeiten. Die Ergebnisse, die im Vortestteil erzielt werden, werden nicht in das Endergebnis eingerechnet, sie dienen ausschließlich der Testung neuer Testitems. Das analytische Schreiben wird immer zuerst, die anderen Testteile in zufälliger Reihenfolge dargeboten, wobei alle Testteile nach einem festen Zeitplan absolviert werden müssen (siehe Tabelle 1).

Tabelle 1, Typischer Zeitplan für einen Papier-Bleistift-GRE (Homepage des GRE)

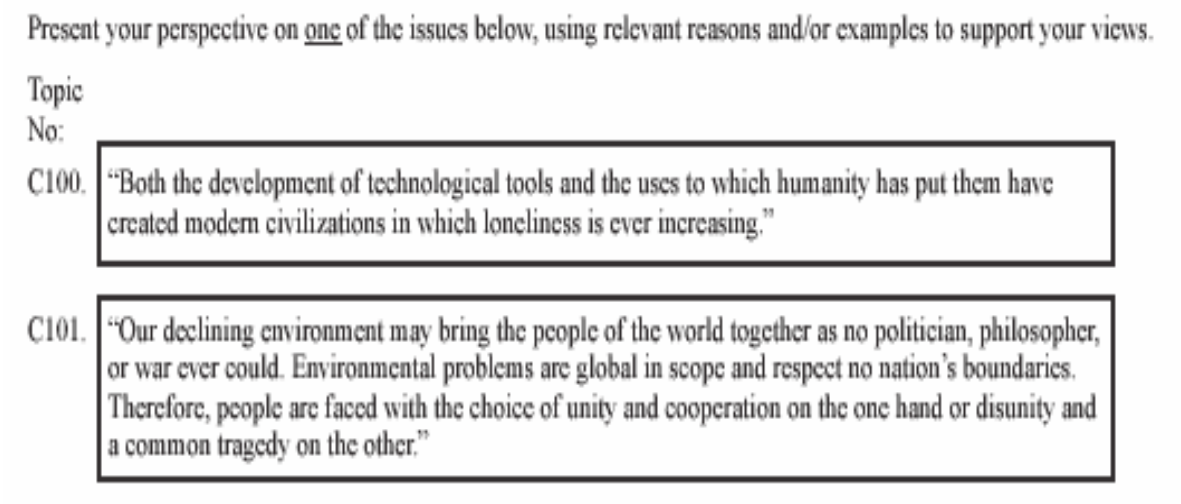
Section	**Number of Questions**	**Time**
Analytical Writing	1 issue task	45 min.
	1argument task	30 min.
Verbal (2 sections)	38 per section	30 min. per section
Quantitative (2 sections)	30 per section	30 min. per section
Pretest	varies	30 min.

Im Folgenden werden die einzelnen Testteile eingehend beschrieben und Beispielaufgaben dazu dargestellt. Die Lösungen zu den Aufgaben des quantitativen sowie des verbalen Testteils befinden sich im Anhang (A.1).

Analytisches Schreiben (Analytical Writing Section)

Im Oktober 2002 wurde ein Testteil, der analytische Testteil, durch einen neuen, der analytisches Schreiben verlangt, ersetzt. Dadurch verspricht man sich eine gesteigerte Testfairness unter den Teilnehmern, die unterschiedliche Voraussetzungen, bezüglich Alter, Geschlecht und Herkunft, mit sich bringen. Außerdem wird eine Verbesserung der Vorhersagevalidität des GRE erwartet. Dieser neue Testteil wurde entwickelt, weil es als wichtig zur Erfassung der Studierfähigkeit der Bewerber erachtet wurde, auch die Fähigkeiten im kritischen und analytischen Denken zu messen. In vielen Studienfächern ist es nämlich ausschlaggebend für den Erfolg im Studium, Argumente formulieren und kritisch diskutieren zu können. Deshalb besteht dieser Testteil aus zwei verschiedenen Aufgabentypen, einer Problemerörterung und einer Diskussion eines Arguments.

Bei der Problemerörterung (Issue Task) hat der Bewerber 45 Minuten Zeit, ein vorgegebenes, recht allgemein gehaltenes, Problem von seinem Standpunkt aus schriftlich, zu erörtern.


Present your perspective on one of the issues below, using relevant reasons and/or examples to support your views.

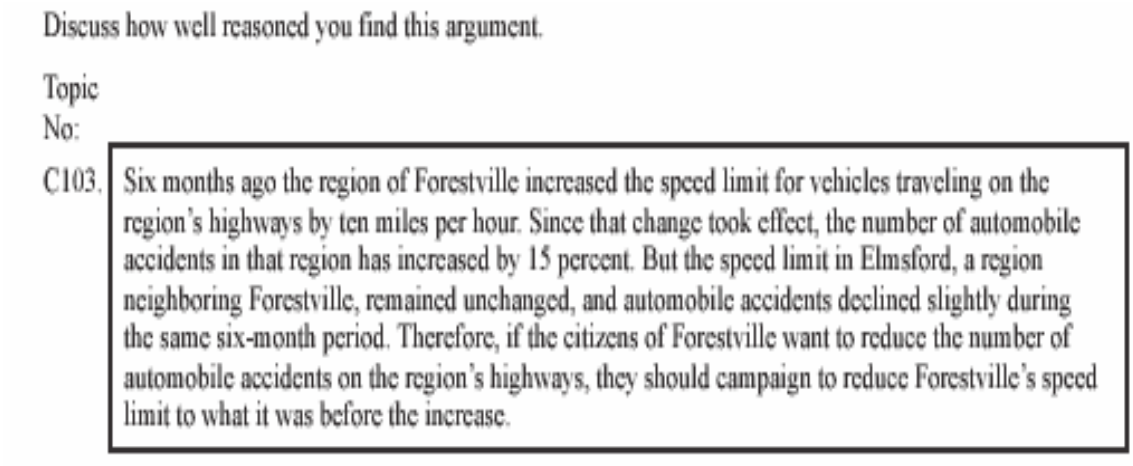
Topic No:

C100. "Both the development of technological tools and the uses to which humanity has put them have created modern civilizations in which loneliness is ever increasing."

C101. "Our declining environment may bring the people of the world together as no politician, philosopher, or war ever could. Environmental problems are global in scope and respect no nation's boundaries. Therefore, people are faced with the choice of unity and cooperation on the one hand or disunity and a common tragedy on the other."

Abbildung 1, Beispielaufgabe 1, Issue Task (Originaltest)

Bei der zweiten Teilaufgabe, der Diskussion eines Arguments (Argument Task), muss der Teilnehmer ein gegebenes Argument zu einem bestimmten Thema auf seine Richtigkeit und Stichhaltigkeit untersuchen. Dazu hat der hat er 30 Minuten Zeit.


Discuss how well reasoned you find this argument.

Topic No:

C103. Six months ago the region of Forestville increased the speed limit for vehicles traveling on the region's highways by ten miles per hour. Since that change took effect, the number of automobile accidents in that region has increased by 15 percent. But the speed limit in Elmsford, a region neighboring Forestville, remained unchanged, and automobile accidents declined slightly during the same six-month period. Therefore, if the citizens of Forestville want to reduce the number of automobile accidents on the region's highways, they should campaign to reduce Forestville's speed limit to what it was before the increase.

Abb. 2, Beispielaufgabe 2, Argument Task (Originaltest)

Quantitativer Testteil (Quantitative Section)

Der quantitative Testteil soll mathematische Basiskenntnisse, das Verständnis elementarer mathematischer Konzepte sowie die Fähigkeit, mathematische Problemstellungen zu lösen, messen. Die dazu nötigen Kenntnisse in Arithmetik, Algebra, Geometrie und Datenanalyse sollen die Bewerber bereits in der High School erworben haben. Im quantitativen Testteil müssen die Teilnehmer Aufgaben aus drei Teilgebieten bearbeiten: Quantitative Vergleiche, quantitatives Problemlösen und Problemlösen durch Dateninterpretation. Soll der Kandidat Quantitative Vergleiche (Quantitative Comparisons) durchführen, muss bei jeder Aufgabe entschieden werden, ob der Ausdruck in der rechten oder in der linken Spalte größer

ist, ob beide gleich sind oder dies nicht entschieden werden kann. Die richtige von den vier Lösungsmöglichkeiten soll vom Kandidaten angekreuzt werden.

Beispielaufgabe 3, Quantitative Vergleiche (Originaltest)

You are to compare the quantity in Column A with the with the quantity

in Column B and decide wether

(A) the quantity in Column A is greater

(B) the quantity in Column B is greater

(C) the two quantities are eaquel

(D) the relationship cannot be determined from the information given

Column A	Column B
9,8	*√100*

Ist der Teilnehmer bei den Aufgaben zu Quantitativem Problemlösen (Problem Solving) angelangt, so ist es die Aufgabe des Kandidaten, allgemeine Mathematik- und Textaufgaben zu lösen. Es werden fünf Lösungsvorschläge angeboten.

Beispielaufgabe 4, Problem Solving (Originaltest)

When walking, a certain person takes 16 complete steps in 10 seconds. At this rate, how many complete steps does the person take in 72 seconds?

(A) 45

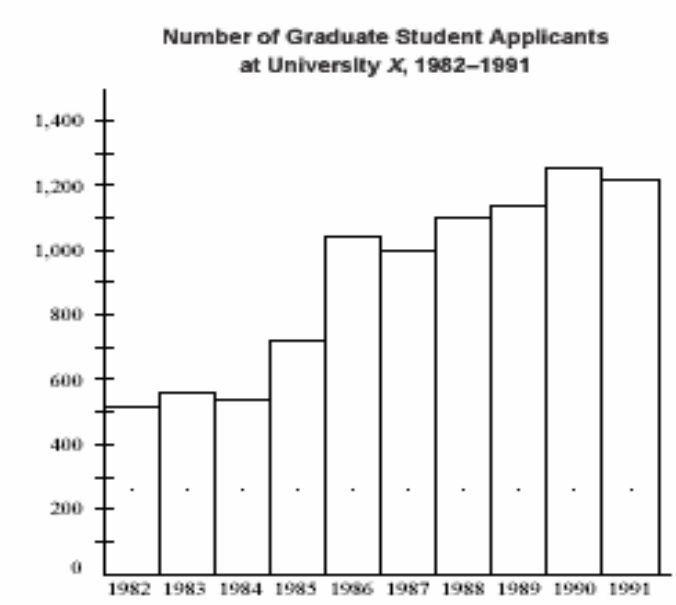
(B) 78

(C) 86

(D) 90

(E) 115

Bei dem Aufgabentyp Problemlösen durch Dateninterpretation (Graph Questions) sollen Fragen zur Interpretation und Auswertung von Graphen beantwortet werden. Der Kandidat kann wieder aus fünf Antwortmöglichkeiten wählen.



In which of the following years did the number of graduate student applicants increase the most from that of the previous year?

(A) 1985
(B) 1986
(C) 1988
(D) 1990
(E) 1991

Abb.3, Beispielaufgabe 5, Graph Questions (Originaltest)

Verbaler Testteil (Verbal Section)

Der verbale Testteil soll die Fähigkeiten der Bewerber, im analysieren schriftlichen Materials, Informationen extrahieren, Beziehung zwischen Satzteilen erkennen und Beziehungen zwischen Wörtern und Konzepten identifizieren, messen. Diese Sektion ist in die vier folgenden Fragetypen unterteilt: Antonyme, Analogien, Satzvervollständigungen und Leseverstehen. Wenn es um Antonyme (Antonyms) geht, soll der Kandidat den Ausdruck bestimmen, der den besten Gegensatz zu einem gegebenen Wort bildet. Dabei kann er aus fünf Antwortmöglichkeiten auswählen. Dieser Fragetyp misst also den Wortschatz und die Fähigkeit Gegensätze zu bilden.

Beispielaufgabe 6, Antonyms (Originaltest)

DIFFUSE:

(A) concentrate

(B) contend

(C) imply

(D) pretend

(E) rebel

Beim Aufgabentyp Analogien (Analogies), ist es die Aufgabe des Testteilnehmers, die richtige Analogie aus fünf Antwortmöglichkeiten auszuwählen. Dabei soll die Fähigkeit, Beziehungen zwischen Wörtern und Konzepten zu erkennen, gemessen werden.

Beispielaufgabe 7, Analogies (Originaltest)

COLOR : SPEKTRUM

(A) tone : scale

(B) sound : waves

(C) verse : poem

(D) dimension : space

(E) cell : organism

Bei dem Fragetyp Satzergänzung (Sentence Completion), muss der Bewerber einen Satz mit zwei Lücken mit den richtigen Wörtern ergänzen. Dabei kann er aus fünf Antwortpaaren wählen.

Beispielaufgabe 8, Sentence Completion (Originaltest)

Early _____of hearing loss is_____by the fact that the other sensesare able to compensate

for moderate amounts of loss, so that people frequently do not know that their hearing is imperfect.

(A) discovery.. indicated

(B) development.. prevented

(C) detection.. complicated

(D) treatment.. facilitated

(E) incidence.. corrected

Das Leseverstehen (Reading comprehension) misst die Fähigkeit, aufmerksam einen Text zu lesen und aus verschiedenen Perspektiven zu analysieren. Die gegebenen Texte stammen aus den Gebieten Geisteswissenschaften, Gesellschaftswissenschaften und Naturwissenschaften. Zuerst muss ein kurzer Text gelesen werden, welcher dem Kandidaten auch während der Beantwortung der Frage vorliegt. Darauf folgen Fragen zum Verständnis, mit fünf Antwortmöglichkeiten, von denen die beste ausgesucht werden soll.

Beispielaufgabe 9, Reading Comprehension (Originaltest)
According to the passage, the two antithetical ideals of photography differ primaly

in the

(A) value that each places on the beauty of the finished product

(B) emphasis that each places on the emotional impact of the finished product

(C) degree of technical knowledge that each requires of the photographer

(D) extent of the power that each requires of the photographer`s equipment

(E) way in which each defines the role of the photographer

Auswertung

Die Ergebnisse der Kandidaten werden beim Papier-Bleistift-Test nach einem festen Lösungsschlüssel von Hand ausgewertet.

Beim GRE im CAT-Format berechnet der Computer selbst das Endergebnis, wobei er nicht nur die richtige Lösung einer Aufgabe, sondern auch ihre Schwierigkeit berücksichtigt. Dies gilt nur für den Verbalen und den quantitativen Testteil. Der verbale Testteil wird nach einer Skala von 200 bis 800, in Zehn-Punkte-Abständen, bewertet. Das gleiche gilt für die quantitative Sektion.

Für das Analytische Schreiben gibt es eine Skala von 0 bis 6, in der die Bewerter Punktwerte in 0,5-Punkte-Abständen vergeben können. Die Bewertung dieser Sektion ist

nicht so einfach wie die der übrigen beiden Testteilen, da kein eindeutiger Lösungsschlüssel vorliegt. Jedes Essay wird von zwei trainierten Lesern bewertet. Sollten sich die Ergebnisse der beiden Bewertungen ummehr als einen Punkt unterscheiden, wird ein drittes Urteil von einem weiteren Leser herangezogen. Ansonsten werden die Punktwerte einfach gemittelt.

Wurde diese Sektion von einem Teilnehmer gar nicht bearbeitet erhält er ein „No Score“. Hat er nur eine der beiden Aufgaben beantwortet, bekommt er in dem unbearbeiteten Teil 0 Punkte. Die Punkte der beiden Aufgaben des Analytischen Schreibens werden gemittelt.

Die Bewerter sollen sich an eingeübte „Scoring“-Richtlinien halten. Das heißt im Wesentlichen, dass sie das meiste Gewicht darauf legen sollen, ob eine Person kritisches Denken und analytische Fähigkeiten erkennen lässt. Grammatik und Rechtschreibung sind hierbei nebensächlich.

Null Punkte sollen vergeben werden, wenn eine Person gar nichts zu einem Thema geschrieben hat. 0,5 bis 1 Punkt erhält der Kandidat, wenn er extrem verworren geschrieben hat, es keine Entwicklung in seinem Text gibt und die angeführten Argumente kaum etwas mit dem vorgegebenen Thema zu tun haben. Der Kandidat bekommt 1,5 bis 2 Punkte, wenn man schwere Mängel in Textentwicklung, Organisation und Satzstruktur, sowie unterdurchschnittlichen Sprachgebrauch feststellen kann. 2,5 bis 3 Punkte kann eine Person erreichen, wenn sie in ihrem Text nur eine schwache Organisation eingebracht hat, ihre Argumente nur vage formuliert sind und eine Entwicklung nur angedeutet ist. Zwischen 3,5 und 4 Punkte erreichen Kandidaten, die komplexe Ideen schon adäquat analysieren können und dabei einige Beispiele zu den Hauptpunkten anführen. Teilnehmer die zwischen 4,5 und 5 Punkte für ihr Essay erhalten, sind in der Lage logisch aufgebaute Analysen zu erarbeiten, mit vielen passenden Beispielen, wobei sie auch auf Satzstruktur und Wortwahl achten. Bekommt eine Person zwischen 5,5 und 6 Punkten, so ist ihr Essay sehr gut strukturiert, logisch konsistent und zielgerichtet. Dazu kommt, dass die Person besonders hohe Fähigkeiten in sprachlichem Ausdruck besitzt (Homepage des GRE). Die drei Einzelergebnisse der Sektionen werden nicht verrechnet, sondern getrennt aufgeführt.

Objektivität

Da es sich beim GRE um ein standardisiertes Testverfahren handelt, kann man die Objektivität insgesamt als sehr hoch einstufen. Um eine gute Durchführungsobjektivität zu gewährleisten, erhält jeder Testteilnehmer die gleiche schriftliche Instruktion und Testzeit und -länge unterscheiden sich nicht zwischen den Teilnehmern. Die Auswertungs-objektivität wird ebenfalls auf einem hohen Standard gehalten. Für den quantitativen, sowie den verbalen Testteil stehen den Auswertern bzw. einem Computerprogramm feste Auswertungsschlüssel zur Verfügung. Bei dem Testteil, in dem die Fähigkeit des Testteilnehmers in analytischem Schreiben untersucht wird, ist es allerdings schwieriger für die Auswerter ein objektives Urteil über die Güte der Arbeit abzugeben. Um letzteres zu

ermöglichen werden die Bewerter des Analytischen Schreibens gut geschult und es wird verlangt, dass neue Bewerter zehn bereits bewertete Texte bearbeiten. Um als Bewerter für den GRE aufgenommen zu werden, müssen ihre Bewertungen zu mindestens 90% mit denen der Vorbewertungen übereinstimmen.

Reliabilität

Zur Messgenauigkeit des GRE liegen nur sehr wenige Studien vor, was sicher mit der Schwierigkeit, die eine Messung der Reliabilität von Leistungstests mit sich bringt, zusammenhängt. Eine Wiederholung des identischen Tests mit der gleichen Person macht auf Grund Erinnerungseffekte nur wenig Sinn.

Zudem müssen die drei verschiedenen Testteile nicht unbedingt besonders hoch miteinander korrelieren, da es durchaus zu erwarten ist, dass eine Person beispielsweise rechnerisch sehr begabt ist, im verbalen Teil hingegen Schwächen aufweist.

Innerhalb einer Sektion sollte die Reliabilität allerdings ausreichend sein.

Für den verbalen Testteil werden auf der Internetseite des GRE für die einzelnen Aufgabentypen Reliabilitäten zwischen .72 und .86 angegeben. Bei dieser Sektion korrelieren die einzelnen Aufgabentypen hoch miteinander (siehe Anhang, A.2).

Validität

Frühere Validitätsstudien mit dem GRE erbrachten äußerst uneindeutige Ergebnisse. Es werden Werte zwischen -.61 und .81 angegeben. In Anschluss an diese Ergebnisse wurde stark daran gezweifelt, ob der GRE weiterhin zur Auswahl der zukünftigen Studierenden eingesetzt werden sollte.

Dies regte weitere Studien und Metaanalysen an. Goldberg und Allinger (1992) fanden eine Validität von .15 bei einer Metaanalyse von insgesamt 10 Studien mit 963 Teilnehmern.

Eine neuere Metaanalyse von Kuncel, Hezlett und Ones (2001), die hier genauer vorgestellt werden soll, beschäftigt sich mit der Vorhersagevalidität des GRE im Bezug auf den Abschlussnotendurchschnitt und verschiedene andere Kriterien.

Diese Metaanalyse zeigt in verschiedenen Aspekten Verbesserungen, verglichen mit früheren Metaanalysen. Sie untersucht die Validitäten für verschiedene Fachrichtungen, wobei die Ergebnisse aus 1521 Studien und davon 6589 Korrelationen analysiert wurden.

Dazu kommt, dass multiple Prädiktoren kombiniert wurden um den Studienerfolg vorherzusagen. Außerdem wurde in anderen Studien nicht direkt auf die statistischen Artefakte, welche die Höhe des Zusammenhangs zwischen GRE-Ergebnis und verschiedenen Kriterien des Studienerfolges beeinflussen, eingegangen.

Oft werden statistische Bedenken gegenüber früheren Validitätsstudien erhoben. Dabei werden meist Begrenzung des Wertebereichs, Kriteriumsunreliabilität und inadäquate Stichprobegröße kritisiert. Der Wertebereich ist deswegen eingeschränkt, weil Validitätsschätzungen des GRE nur mit Personen durchgeführt werden können, welche die Zulassung zum Studium erhalten haben, da als Kriterium für den Studienerfolg oft beispielsweise die Klausurnoten herangezogen wird. Da viele Hochschulen sich bei der Studierendenauswahl sehr stark an den GRE-Werten orientieren, ist die Varianz der Ergebnisse unter allen Bewerbern größer, als die unter den angenommenen Kandidaten. Dadurch wird die Validität des GRE im Allgemeinen stark unterschätzt. Außerdem sind die Messungen des Studienerfolges meist unreliabel und diese Messfehler beeinflussen die Vorhersagevalidität des GRE.

Diese Metaanalyse korrigiert den Einfluss der Messwertebeschränkung und Kriteriumsunreliabilität, indem das Kriterium, also der Studienerfolg, multidimensional definiert wird. Dabei werden die folgenden acht Kriterien berücksichtigt: Note im Bachelorabschluss, Noten im ersten Studienjahr, Klausurnoten, Bewertungen der Fakultäten, Anzahl der Publikationen, Anzahl, wie oft die Publikationen zitiert wurden, Titelerwerb und Zeit bis zum Erwerb eines Titels.

Bachelor-Noten und Noten im ersten Jahr

Die Kriterien, Bachelor-Note und Noten im ersten Studienjahr werden am häufigsten herangezogen um den Studienerfolg zu messen. Dies hat den Vorteil, dass Arbeitsverhalten über einen längeren Zeitraum, Wissenserwerb, Anstrengung und Fähigkeit gemessen werden. Nachteilig ist allerdings die Tatsache, dass es erhebliche Unterschiede zwischen den verschiedenen Universitäten, Landkreisen und sogar im gleichen Fach zwischen verschiedenen Professoren gibt. Man erwartet mittlere Korrelationen zwischen GRE-Ergebnis und Noten im Bachelorabschluss, bzw. den Noten im ersten Studienjahr. Der GRE Subject Test dürfte in engerem Zusammenhang damit stehen.

Klausurnoten

Die Klausurnoten zeigen an, wie gut die Studenten den Lernstoff verstanden haben und wie weit sie in der Materie vorangeschritten sind. Allerdings können sich Klausuren hinsichtlich ihrer Schwierigkeit, ihrer Relevanz für das Studium und in verschiedenen Studienabschnitten unterscheiden. Hier werden ähnliche Korrelationen mit dem GRE-Ergebnis erwartet, wie bei den Bachelor-Noten.

Bewertungen durch den Fachbereich

In die vorgestellte Metaanalyse, wurden nur solche Studien eingeschlossen, in denen sich die Bewertungen auf allgemeines Können, Praktika und Forschungsarbeit bezogen. Bei multiplen Validitäten, wurde für jeden Prädiktor der Durchschnitt genommen. Bewertungen

haben den Vorteil, flexibel zu sein und viele Bereiche abdecken zu können. Allerdings können sie abhängig von den Vorlieben des Bewerters sein und von anderen Fehlern, wie dem Halo-Effekt und der Zentralen Tendenz abhängen. Es werden mittlere bis hohe Korrelationen mit dem GRE erwartet.

Forschungsproduktivität

Die Forschungsproduktivität wird meistens durch die Anzahl der Publikationen oder anderer wissenschaftlicher Schriftstücke, welche die Person in ihrer Studienzeit oder danach veröffentlicht hat, ausgedrückt. Wissenschaftliche Produktivität ist das Ziel der meisten Studiengänge. Trotzdem planen nicht alle Studenten eine wissenschaftliche Karriere. Dazu kommt, dass Quantität nicht unbedingt mit Qualität einhergehen muss. Erwartet werden demnach niedrige aber positive Korrelationen mit dem GRE. Eng damit verbunden ist das Kriterium wie oft die Publikationen einer Person in anderen Publikationen zitiert werden. Da man im Allgemeinen davon ausgehen kann, dass Arbeiten hoher Qualität öfter zitiert werden, ist die Anzahl der Zitate ebenfalls ein Ausdruck für Studienerfolg.

Titelerwerb und Zeit bis zum Titelerwerb

Der Titelerwerb hängt von einem weiten Spektrum an Umständen ab. Diese können die schulische Aktivität oder auch interpersonelle Beziehungen des Kandidaten widerspiegeln. Nur einige der Unterschiede zwischen verschiedenen Personen betreffen die Fähigkeit selbst. Außerdem wollen nicht alle Personen einen Titel erwerben. Die Autoren der Metaanalyse erwarten geringe, aber positive Korrelationen des Titelerwerbs mit dem GRE-Ergebnis einer Person.

Potentielle Moderatoren der Validitäten des GRE

Manche Variablen können den Zusammenhang zwischen GRE und Studienerfolg beeinflussen. Die Validität des GRE ist möglicherweise abhängig vom Studienfach des Kandidaten. Es gibt zwar einige Gemeinsamkeiten, aber auch viele Unterschiede hinsichtlich der Anforderungen und Lernprogramme verschiedener Studiengänge. Um diese Moderatorvariable zu erfassen, wurden separate Analysen für die vier Fachrichtungen: Geisteswissenschaften, Gesellschaftswissenschaften, Umweltwissenschaften und mathematisch-physikalische Wissenschaften gemacht. Eine zweite Moderatorvariable könnte sein, ob die Muttersprache der Kandidaten Englisch ist oder nicht. Bei den verbalen Testteilen könnten daher die Nicht-Muttersprachler etwas im Nachteil sein. Die Validität für den quantitativen Testteil wird demnach etwas höher vermutet. Außerdem könnte das Alter der Studenten eine moderierende Funktion haben. Ältere Studenten unterscheiden sich von jüngeren in der Zeit die sie an der Universität verbringen, in der Lebenserfahrung und in persönlichen Gebundenheiten. Trotzdem werden die Validitäten des GRE für alle Altersgruppe etwa gleich hoch erwartet.

Methode

Insgesamt behandelt die Metaanalyse die folgenden drei Fragestellungen: Wie valide ist der GRE als Prädiktor für Studienerfolg? Ist der GRE ein besserer Prädiktor für verschiedene Kriterien als andere? Haben Moderatorvariablen einen Einfluss auf die Validität des GRE? Um diese Fragestellungen zu beantworten, wurden die gesammelten Daten mit der psychometrischen Metaanalysemethode von Hunter und Schmidt (1990) analysiert. Damit ist es möglich, die Varianz, die durch Stichprobenfehler, Messwertebeschränkung und Mängel bei der Reliabilität zustande kommt, zu identifizieren. Es wurde eine interaktive Prozedur der Metaanalyse angewandt.

Beschreibung der Datenbasis

Verschiedene Studien, Dissertationen und technische Berichte des ETS, die Studienerfolg und den GRE untersuchen wurden in die Metaanalyse eingeschlossen. Die Informationen, die aus den Berichten extrahiert wurden waren: Art des Prädiktors, Art des Kriteriums, Effektgröße, Stichprobengröße, Moderatorvariablen, mögliche Einschränkung des Wertebereichs und Kriteriumsunreliabilität. Es gab bis zu 29 Informationen pro bivariatem Zusammenhang. Bei Überlappung von Stichproben, wurden nur die Studien eingeschlossen, die umfassender oder kompletter waren. Studien, die nur die signifikanten Ergebnisse berichteten, wurden nicht in die Metaanalyse einbezogen. Die gesamte Datenbasis beträgt 1 753 unabhängige Stichproben, 6 589 Korrelationen und 82 659 Studenten.

Allgemeine Ergebnisse

Zuerst hat man die Validitäten der Bachelor-Abschlussnote und des GRE zur Vorhersage der Abschlussnote über alle Studien ermittelt.

Man konnte mittlere Validitäten für den verbalen Testteil des GRE (GRE-V) (N=14 156, k=103), für den quantitativen Testteil (GRE-Q) (N=14 425, k=103), für den analytischen Testteil (GRE-A) (1 928, k=20) und für die Bachelor-Noten (N=9 748, k=58) feststellen. Es resultierte folgende Validitäten: GRE-V: [rho]= .34, GRE-Q: [rho]=.32, GRE-A: [rho]= .36 und Bachelor-Noten: [rho]= .30. Die Standardabweichungen waren sehr gering, woraus man schließen kann, dass Moderatorvariablen einen äußerst geringen Einfluss ausüben. Das Konfidenzintervall schließt die Null nicht mit ein, was bedeutet, dass es sich bei den GRE-Skalen und den Vordiplomsnoten um valide Prädiktoren handelt. Die Subject Tests des GRE hatten sogar noch höhere Validitäten ([rho]= .41) und geringere Standardabweichungen.

Die Ergebnisse zur Vorhersage der Noten im ersten Studienjahr waren denen der Abschlussnoten sehr ähnlich. Dabei war die Validität der Subject Tests noch etwas höher: [rho]= .45. Sollten die Klausurnoten vorhergesagt werden, resultierte für den GRE-V (N=1198, k=11) eine Validität von [rho]= .44, für den GRE-Q (N=1194, k=11) eine

Validität von [rho]= .26, für die Noten im Bachelorabschluss (N= 592,k=6) eine Validität von [rho]= .12 und für die Subject Tests (N= 534, K=4) eine Validität von [rho]= .51. Die Standardabweichungen gleichen denen, für die Abschlussnoten.

Die Validitäten für die Vorhersage von Beurteilungen der Fakultät sind für den GRE-V (N= 4766, k= 35) [rho]=.42, für den GRE-Q (N=5112, k= 34) [rho]= .47, für den GRE-A (N= 1982, k= 9) [rho]= .35, für die Bachelor-Noten (N= 3695, k= 22) [rho]= .35 und für die Subject Tests (N= 879, K= 12) [rho]= .50.

Möchte man den Titelerwerb vorhersagen, resultieren für den GRE-V (N= 6304, k= 32), den GRE-Q (N= 6304, k=32), den GRE-A (N= 1233, k=16) und die Bachelor-Noten (N= 6315, k= 33) Validitäten von [rho]= .11 bis .20. Die Subject Tests (N= 2575, k= 11) haben hier eine höhere Validität von [rho]= .39. Es werden höhere Standardabweichungen festgestellt, was auf den Einfluss von Moderatorvariablen schließen lässt.

Die Validitäten für die Zeit bis zum Titelerwerb sind folgende: GRE-V (N= 130, k= 3) [rho]= .28, GRE-Q (N= 160, k=3) [rho]= -.12, Bachelor-Noten (N=629, k= 5) [rho]= -.08 und Subject Tests (N= 66, k= 2) [rho]= .02.

Für die Forschungsproduktivität konnte man für den GRE-V (N= 3328, k= 18) und für den GRE-Q (N= 3328, k= 18) leicht positive Validitäten ermitteln. Die Subject Tests (N= 3058, k= 16) weisen eine Validität von [rho]= .21 auf. Der Zusammenhang mit der Zitatanzahl liegt für den GRE-V (N= 2306, k= 12) bei [rho]= .23, für den GRE-Q (N= 2306, k= 12) bei [rho]= .17 und für die Subject Tests (N= 2306, k= 12) bei [rho]= .24 (siehe Anhang A.3, Abb. A 1).

Ergebnisse aus verschiedenen Studiengebieten

Es wurde zwischen vier verschiedene Studiengebiete unterschieden. Zu den Geisteswissenschaften gehörten beispielsweise die Fächer: Kunst, Musik und Philosophie, zu den Gesellschaftswissenschaften: Psychologie, Pädagogik und Politikwissenschaften, zu den Umweltwissenschaften: Biologie, Landwirtschaft und Tiermedizin und zu den Mathematisch-Physikalischen Wissenschaften: Mathematik, Physik und Chemie. Die Ergebnisse für die separaten Studiengänge waren untereinander und mit den Gesamtergebnissen sehr ähnlich. Wiederum waren die Validitäten der Subject Tests bei allen Studienrichtungen höher. Eine Ausnahme stellte wieder der Titelerwerb dar. Es liegen geringe Standardabweichungen vor.

Ergebnisse für Nicht-Muttersprachler und ältere Studenten

Die Studien, die Nicht-Mutersprachler betrafen, wurden nicht in die ganzheitliche Analyse miteinbezogen. Der GRE-V(N= 1764, k=6) sagte bei Nicht-Muttersprachlern die Abschlussnoten nicht so gut voraus ([rho]= .36) wie der GRE-Q (N= 1705, k= 5, [rho]= .53). Der Notendurchschnitt im ersten Studienjahr wurde ebenfalls durch den GRE-Q

([rho]= .40) besser vorhergesagt als durch den GRE-V ([rho= .22]) und durch den GRE-A ([rho]= .35). Für den Erfolg im Studium von Studenten die älter als 30 Jahre sind ist der GRE mit positiven, aber nicht sehr hohen Validitäten, auch ein valider Prädiktor.

Ergebnisse für kombinierte Prädiktoren

Um die Validitäten kombinierter Prädiktoren zu ermitteln wurden Interkorrelationsmatrizen für verschiedene Prädiktoren erstellt und auch Interkorrelationen für die unterschiedlichen Kriterien wurden geschätzt. Dabei stellte sich heraus, dass es von Vorteil ist die beiden Kriterien Abschlussnotendurchschnitt und Fachbereichsbeurteilung zu kombinieren, weil sie beide gut erforscht sind und wichtige Aspekte des Studienerfolges repräsentieren. GRE-V, GRE-Q und Bachelor-Noten kombiniert, sagen das kombinierte Kriterium sogar besser voraus ([rho]= .53) als die Subject Tests alleine ([rho]= .49).

Diskussion

Zusammenfassend kann man also sagen, dass der GRE-V, der GRE-Q, der GRE-A und die Subject Tests valide Prädiktoren für Abschlussnote, Noten im ersten Studienjahr, Beurteilungen durch die Fakultät, Klausurnoten, Zitatanzahl und nicht valide zur Vorhersage von Titelerwerb sind.

Die insgesamt geringen Standardabweichungen signalisieren, dass die Ergebnisse nicht besonders stark von Moderatorvariablen beeinflusst werden.

Die GRE Subject Tests waren bessere Prädiktoren für die meisten Kriterien und waren außerdem gute Prädiktoren für Forschungsproduktivität. GRE-V, GRE-Q, GRE-A und Noten im Bacheloreabschluss lieferten für sämtliche Kriterien etwa gleichgute Vorhersagen.

Auch Variablen wie Studienrichtung, Muttersprache und Alter beeinflussen die Validität des GRE nicht merklich. Es gibt allerdings Faktoren, welche die Validität verfälschen könnten. Den Universitäten ist das GRE-Ergebnis des Studierenden bekannt. Dieses Wissen könnte die Notenvergabe sowie die Beurteilung beeinflussen. Normalerweise sind aber die GRE-Ergebnisse der Studenten, nach der Zulassung, der Fakultät für die Notenvergabe nicht mehr zugänglich.

Der GRE erweist sich als unzureichender Prädiktor für den Titelerwerb, was man mit dem starken Einfluss nichtkognitiver und situativer Faktoren erklären könnte. Die Tatsache, dass die Subject Tests hier bessere Ergebnisse erbringen, hängt damit zusammen, dass hohe Subject Test-Werte, sowie Titelerwerb mit der Höhe des Interesses am Studienfach zusammenhängen.

Kritik

Das Ziel der Auswahlverfahren der Hochschulen ist es nur die besten Bewerber, die ihr Studium am wahrscheinlichsten erfolgreich abschließen werden, zum Studium zuzulassen. Oft wird das GRE-Ergebnis als gewichtiges Kriterium eingesetzt. Daher muss man sich auch kritisch mit diesem Eignungstest auseinandersetzen, um herauszufinden ob sich das Vertrauen, das dem GRE entgegengebracht wird, auch wirklich lohnt. Außerdem kostet die Vorbereitung für den Test und der Test selbst die angehenden Studenten viel Zeit und Geld. Ein Kritikpunkt am GRE ist, dass es sehr viel mehr Faktoren gibt die den Studienerfolg beeinflussen, als im GRE berücksichtigt werden (siehe Abb.4). Die Basis des Tests sind verbale, quantitative und analytische Fähigkeiten, während kreative und praktische kaum einbezogen werden. Allerdings wurde der GRE in dieser Hinsicht durch die Aufnahme der Sektion zum Analytischen Schreiben schon stark verbessert.

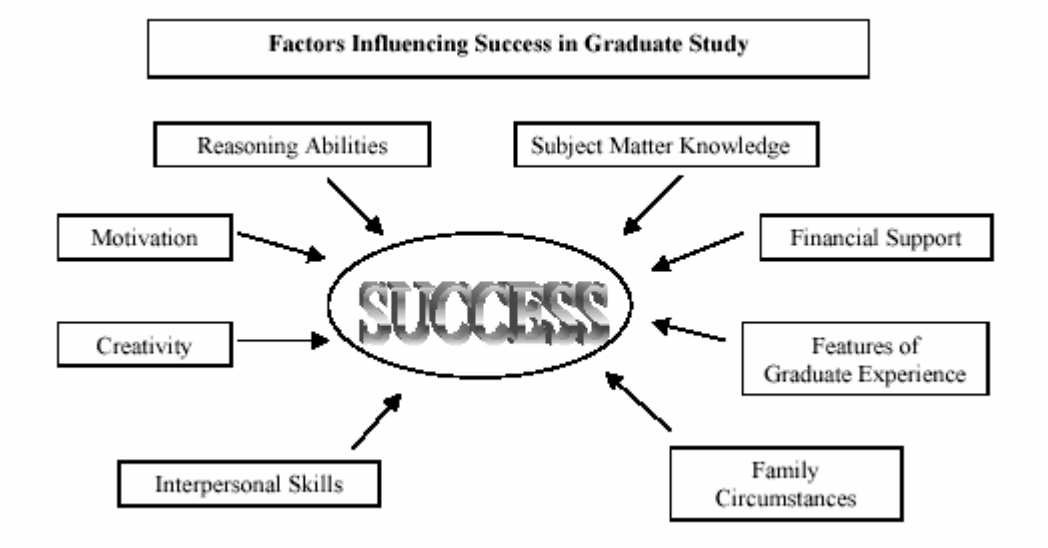


Abb.4, Faktoren, die den Studienerfolg beeinflussen (GRE-Homepage)

Eine Studie von R. Sternberg und W. Williams (1997) hat sich mit den Problemen des GRE als Mittel zur Studierendenauswahl beschäftigt. Ihre Forschung beschränkte sich auf 170 Psychologiestudenten an der Universität von Yale.

Sternberg und Williams stellen heraus, dass der GRE für einige Kriterien gute, für andere nur unzureichende Validitäten aufweist. Es stellt sich die Frage ob es sich bei den Kriterien für die der GRE ein guter Prädiktor ist auch um die relevantesten handelt. Die Autoren sehen das Problem bei der Forschung mit dem GRE, darin dass es, wenn für einen bestimmten Studiengang nur Personen zugelassen werden, die einen hohen GRE-Wert erreicht haben, unmöglich wird die Ansicht, dass hohe GRE-Werte für den Studienerfolg essentiell sind, zu falsifizieren. Zudem bekommen Studenten die ein sehr gutes Ergebnis im GRE haben eher ein Stipendium, was ihnen in der Regel eine Hilfe bei der Organisation des Studiums ist.

Sternberg und Williams beziehen sich auf die Ergebnisse früherer Studien, besonders auf das Kapitel IX des GRE Technical Manual (Briel et al., 1993). Darin wurden drei Faktoren für Studienerfolg gefunden, die sehr gut zu den drei Testteilen des GRE passen. Aber diese Faktoren korrelieren sehr stark untereinander. Verbal und Quantitativ korrelieren zu .64, Verbal und Analytisch zu .77 und quantitativ und Analytisch ebenfalls zu .77. Außerdem fand man in früheren Studien, dass die Validitäten des GRE sich abhängig vom Teilbereich voneinander unterscheiden. Auch haben die Subject Tests, wie oben bereits erwähnt, meist eine höhere Validität.

Ein weiteres Problem auf das Sternberg und Williams hinweisen ist, dass sich theoretische Problemsituationen oft von praktischen, im echten Leben, unterscheiden. Standardisierte Tests wie der GRE können demnach den Erfolg in realen Situationen nicht gut vorhersagen. Daher sagt der GRE auch nur die Noten im ersten Studienjahr nach Sternberg und Williams Befunden einigermaßen adäquat voraus. Später im Studium kommt es nämlich eher auf praktische Fähigkeiten, wie zum Beispiel Referate vorbereiten und Untersuchungen planen, an. Ob die Noten im ersten Jahr ein Auswahlkriterium darstellen sollten, ist allerdings fraglich. Man sollte viel eher vorhersagen können, ob beispielsweise ein Psychologiestudent später ein guter Psychologe sein wird. Man könnte die bessere Vorhersage des GRE im Bezug auf die Noten im ersten Studienjahr auch lediglich damit erklären, dass das erste Jahr zeitlich näher an der GRE-Testung liegt. Somit haben sich die Interessen und die Motivation noch nicht sehr stark in eine andere Richtung bewegen können wie nach zwei oder mehr Jahren.

Die Autoren schlagen vor, die Studierendenauswahl basierend auf der Triarchic Theory of Human Intelligence (Sternberg, 1985), zu verbessern. Sternberg unterscheidet in seiner Theorie zwischen wissenschaftlich-analytischen, synthetisch-kreativen und praktisch-kontextualen Aspekten menschlicher Fähigkeit. Diese drei Faktoren korrelieren nur sehr schwach miteinander. Nach dieser Theorie wäre nur einer dieser Faktoren entscheidend für das Ergebnis beim GRE, nämlich der wissenschaftlich-analytische. Sternberg (1996) fand aber, dass es zur Vorhersage des Erfolges in Kursen bei denen es auch auf praktische und kreative Fähigkeiten ankommt, auch wichtig ist, diese Aspekte in die vorangehende Eignungstestung einzuschließen.

In ihrer Studie in Yale fanden Sternberg und Williams lediglich bescheidene Zusammenhänge zwischen GRE-Ergebnis und Studienerfolg. Eine etwas bessere Validität wurde bei den Noten im ersten Studienjahr erreicht. Diese lag allerdings auch nur bei .17. Im zweiten Studienjahr war sie aber noch geringer und lag bei .02. Die Subject Tests erwiesen sich mit einer Validität von .37 als valide Prädiktoren für die Noten im ersten Studienjahr.

Die Tatsache, dass die Universität von Yale ein sehr strenges Auswahlverfahren hat, relativiert aber Sternbergs Ergebnisse. Die strenge Auswahl führt nämlich zu einer Begrenzung des Wertebereichs für GRE-Ergebnisse, Notendurchschnitte und Qualität der

Dissertationen (Thomas und Helgland, 1998). Da die Universität von Yale einen sehr guten Ruf hat, hat die Selbstselektion der Studenten ebenfalls einen starken Einfluss. Bei Sternbergs Stichprobe handelt es sich Personen mit sehr hohen GRE-Werten. Zum Beispiel ist der Mittelwert der Yale Studenten im verbalen Testteil des GRE um 1,5 Standardabweichungen höher als der Populationsmittelwert (Roznowski, 1998). Was zusätzlich die Validität des GRE in dieser Studie reduzieren könnte, ist die Tatsache, dass nicht nur nach dem GRE entschieden wird ob ein Kandidat zugelassen wird, sondern auch nach kompensatorischen Auswahlregeln. Dies bedeutet, dass auch Studenten mit geringeren Werten im GRE hin und wieder angenommen werden, wenn sie die schlechten Ergebnisse mit ausgezeichneten Praktikumszeugnissen beispielsweise ausgleichen können. Außerdem kann man die Stichprobengröße als zu gering einschätzen. Dies führt ebenfalls zu einer Unterschätzung der Validität des GRE.

Sternberg fordert nach seinen Ergebnissen einen erweiterten, verfeinerten Test zur Studierendenauswahl und empfiehlt seinen Triarchic Abilities Test. Eine Verbesserung ist durchaus ein erstrebenswertes Ziel. Trotzdem kann man den GRE nicht als nutzloses Instrument bezeichnen. Wenn Sternberg und Williams ihre Hypothese, dass der GRE kein guter Prädiktor für Studienerfolg ist, ausreichend belegen wollen, müssten sie eine adäquate Methode anwenden, bei der ihre Hypothesen auch falsifizierbar sind (Melchert, 1998). Die Yale-Stichprobe war dazu denkbar schlecht gewählt.

Man kann den GRE also, besonders nach den Ergebnissen der Metaanalyse von Kuncel et al. (2001), als validen Prädiktor für Studienerfolg bezeichnen. Dies schließt natürlich nicht aus, dass verschiedene Verbesserungen an ihm vorgenommen werden könnten. Vielleicht wäre es aber auch sinnvoll sich zunächst einmal den Kriterien zuzuwenden und eine Methode zu finden Notenvergabe, Beurteilungen und dergleichen reliabler zu machen. Es ist zu erwarten, dass damit auch die Validität des GRE ansteigt.

Zusammenfassung

Der GRE, der als Instrument zur Studierendenauswahl häufig von Hochschulen in den USA eingesetzt wird, hat, wie man erkennen kann, seine Daseinsberechtigung. Dass seine ermittelten Validitäten nicht allzu hoch sind kann man nicht nur auf die Unzulänglichkeit des GRE, Studienerfolg vorherzusagen zurückführen, sondern auch auf verschiedene andere Faktoren, wie die mangelhafte Reliabilität der Kriterien. Trotzdem kann der GRE noch optimiert werden und mit der Einführung der Sektion des Analytischen Schreibens hat man schon eine deutliche Verbesserung erzielt. Der GRE ist sicherlich ein Test, der in ähnlicher Form, auch in Deutschland als Teil eines Studierendenauswahlverfahrens eingesetzt werden könnte.

Literatur

Briel, J. B., O'Neil, K. & Scheunemann, J. D. (Eds.). (1993). GRE technical manual. Princeton, NJ: Educational Testing Service.

Goldberg, E. L., & Allinger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalisation approach. Educational and Psychological mesurement, 52, 1019-1027.

Kuncel, N. R., Hezlet, S. A. & Ones, D. (2001). A Comprehensive Meta-Analysisof the Predictive Validity of the Graduate Record Examinations: Implicationsfor Graduate Students Selection and Performance. Psychological Bulletin, 127, 162-181

Melchert, T. (1998). Support for the validity of the Graduate Record Examination. Texas Tech University.

Moosbrugger, H. (1997). Testmodelle der Item-Response-Theorie (IRT). Johann Wolfgang Goethe Universität Frankfurt am Main.

Roznowski, M. (1998). The Graduate Record Examination: Testing :: Yale University : Academics [Comment]

Sternberg, R. J., (1985). Beyond IQ: A triarchic thoery of human intelligence. New York. Cambridge University Press.

Sternberg, R. J, (1996). Succesful intelligence. New York. Simon & Schuster.

Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful succes in the graduate training of psychologists? A case study. American Psychologist, 52, 630-642. Ovid full Text Library Holdings Accession Number: 00000487-199805000-00012

Thomas, A. & Helgland, S., (1998). Range Restriction, Outliers, and the Use of the Graduate Record Examination to Predict Graduate School Performance. Iowa State University.

WWW.GRE.ORG [Stand: 20.02.2005]

Anhang

A.1 Lösungen zu den Beispielaufgaben:

Beispielaufgabe 3: (B)

Beispielaufgabe 4: (E)

Beispielaufgabe 5: (B)

Beispielaufgabe 6: (A)

Beispielaufgabe 7: (A)

Beispielaufgabe 8: (C)

Beispielaufgabe 9: (E)

A.2 Reliabilität

Correlations, Reliabilities (in diagonal), and Disattenuated Correlations for GRE Verbal Subtests, Form 3CGR1

	Reading Comprehension	Sentence Completion	Antonym	Analogy
Reading Comprehension	(.790)	.899*	.768*	.847*
Sentence Completion	.677	(.718)	.863*	.894*
Antonym	.632	.677	(.858)	.909*
Analogy	.649	.653	.726	(.743)

A.3 Validität

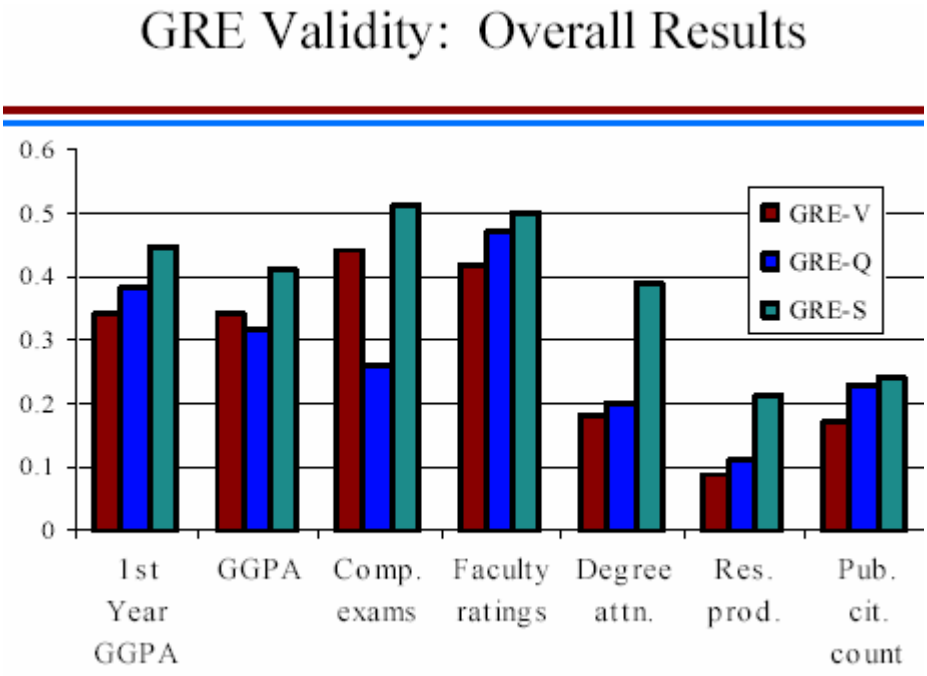


Abb. A 1, Allgemeine Ergebnisse