Zentrum für Testentwicklung und Diagnostik
am Psychologischen Institut der Universität Fribourg

Eignungsprüfung für das Medizinstudium

Kriterien und Testverfahren

Bericht über das Internationale Symposium in Bern
8. November 1996
herausgegeben von K.-D. Hänsgen und N. Ischi

Bericht 3 (1996)

Inhaltsverzeichnis

Redaktion: Judith Berger

am Psychologischen Institut der Universität Fribourg
Route d'Englisberg 9, CH-1763 Granges-Paccot
Tel. 026/300 79 86 - Fax 026/300 97 63 - e-mail: ZTD@UNIFR.CH
Internet: http://www.unifr.ch/Main/ztd.html

Vorwort der Herausgeber

In allen hochentwickelten Industrieländern der Welt geht der Trend hin zu einer stärkeren Nachfrage nach Hochschulausbildung, die auch vor dem Medizinstudium nicht Halt macht. In vielen Ländern ist im Fach Medizin ein Numerus Clausus als Beschränkung der Zulassung entsprechend der Ausbildungskapazitäten der Universität gängige Praxis, um die Arbeitsfähigkeit der Universitäten und die Qualität der Ausbildung aufrecht zu erhalten. Beschränkungen machen in jedem Falle sowohl genügende gesetzliche Grundlagen als auch ein wissenschaftlich gesichertes, faires, und anerkanntes Kriterium notwendig, nach welchem die Zulassung erfolgen kann.

Bekanntlich ist in der Schweiz seit mehreren Jahren die Tendenz vorhanden, dem grossen Interesse am Medizinstudium unter der Bedingung eines bisher unbegrenzten Hochschulzugangs durch Überschreitung der Ausbildungskapazitäten zu begegnen. Rund ein Drittel der Zugelassenen beenden ihr Medizinstudium vorzeitig während der ersten zwei Studienjahre - teilweise als Folge einer verstärkten intrauniversitären Selektion, die als “Notmassnahme” eingesetzt werden muss, um die Zahl der Studierenden den vorhandenen Kapazitäten in der klinischen Ausbildung anzupassen. Diese Kapazität wird unter anderem entscheidend begrenzt durch die Patienten- bzw. Bettenzahlen in den Spitälern.

Die Schweiz gehört bekanntlich zu den Ländern mit der höchsten Dichte an Studienplätzen im Fach Medizin und somit mit der höchsten Ärztedichte. Vor dem Hintergrund, dass der bereits ausserordentlich grosse Anteil der finanziellen Ressourcen, welcher zugunsten der medizinischen Fakultäten aufgewendet wird, nicht mehr erhöht werden kann, generell einer Kostenexplosion im Gesundheitsbereich rechtzeitig Einhalt geboten werden muss und auch die Mittel der Universitäten rationell einzusetzen sind, wird von politischer und akademischer Seite ein Numerus Clausus als Notmassnahme immer wieder in Erwägung gezogen. Nicht zuletzt wird auch das Absinken der Ausbildungsqualität von den Universitäten und den Studierenden selbst beklagt und die Notwendigkeit und Dringlichkeit von Studienreformen im Fach Medizin anerkannt. Die an den Universitäten Genf, Lausanne und Bern bereits eingeleiteten Studienreformen erfordern jedoch eine zusätzliche Reduktion der Aufnahmekapazitäten.

Bedenkt man, dass die Kantone und der Bund für die Ausbildung jedes Studierenden im Fach Medizin in den ersten beiden Studienjahren im Durchschnitt rund 60'000 Franken pro Jahr aufwenden müssen, so ist heute die hohe Zahl der Studienabbrüche auch aus wirtschaftlichen Gründen nicht mehr zu vertreten. Wenn das **Kriterium der Studieneignung** verwendet wird und die Zulassung nach Massgabe der Kapazität erfolgt, ist die Wahrscheinlichkeit einer vorzeiti-

gen Beendigung des Studiums wesentlich geringer und die getätigten Investitionen führen dann auch zu tatsächliche Leistungen.

Im Jahre 1995 sind erstmals gesamtschweizerisch Vorbereitungen für den Einsatz eines Eignungstests als Zulassungskriterium in Angriff genommen worden. Verschiedenste Gründe - u.a. fehlende gesetzliche Grundlagen und Umstellungen des Schuljahres bei einigen Gymnasien - hatten zur Folge, dass der Einsatz des Eignungstests nochmals hinausgeschoben werden konnte. Das Überlastungsproblem besteht allerdings weiter und es gibt keine Aussicht auf Selbstlösung. Zudem zeichnet sich ab, dass in einigen Kantonen neuerdings eine alternative Lösung zum Eignungstest bevorzugt wird, namentlich die verstärkte Selektion im ersten Studienjahr. Man muss sich dabei allerdings der Tatsache bewusst sein, dass auf hochschulpolitisch sinnvolle und stabile Lösungen langfristig weiterhin nur gesamtschweizerisch hingearbeitet werden kann. Schon im vorigen, sicher aber in diesem Jahr wird es für die Bewerberinnen und Bewerber nicht einfacher, bei der Wahl der Universität, an der sie Medizin studieren möchten, die Vor- und Nachteile der verschiedenen Studienbedingungen genau abzuschätzen.

Weil sich das Problem nur gesamtschweizerisch lösen lässt, ist das Kennenlernen international bewährter Lösungen zur Regulierung der Zulassung wichtig. In vielen Ländern liegen zum Teil langjährige Erfahrungen auf diesem Gebiet vor, die es zu nutzen gilt. Aus diesem Grunde wurde am 8. November 1996 in Bern ein internationales Symposium zum Thema “Eignungsprüfung für das Medizinstudium - Kriterien und Testverfahren” durchgeführt.

Beiträge von acht eingeladenen Referenten wurden von den Vertretern der Gymnasien, der Universitäten, der Universitätskliniken, der akademischen Studienberatungen, der medizinischen Fachverbände sowie der hochschul- und gesundheitspolitischen Organe sehr angeregt diskutiert:

G. Trost stellte die **Ergebnisse einer Umfrage** vor, die ermittelte, welche Massnahmen in welchen **Ländern Europas** bei der Zulassung zum Medizinstudium angewandt werden. Es gibt kaum noch Länder, die sich den Luxus leisten können, nichts zu tun. Standardisierte Tests kommen dabei sehr häufig zum Einsatz. An die Kriterien dieser Testverfahren sind jeweils hohe Anforderungen zu stellen.

Im vorliegenden Symposiumsbericht werden zwei praktizierte Testkonzepte erläutert: J. Hackett stellt den **MCAT (Medical College Admission Test)** vor, der in den USA und in Kanada zum Einsatz gelangt. I. Wedman berichtet in seinem Beitrag über den schwedischen Test SweSAT. Beide Tests können neben dem deutschen Test für medizinische Studiengänge (TMS) als wissenschaftlich gut gesicherte Instrumente gelten, die in den jeweiligen Ländern auch die notwendige Akzeptanz finden.

P.J. Janssen legt in seinem Beitrag Überlegungen zur **Einführung eines Tests in Belgien** dar - einem Land, welches beispielsweise wegen seiner Mehrsprachigkeit mit der Schweiz vergleichbare Rahmenbedingungen hat.

In den Beiträgen von U. Schallberger und K.-D. Hänsgen wird auf die Situation der Schweiz eingegangen. Es werden die praktisch vollzogenen Schritte vorgestellt und die Anforderungen an ein Zulassungsverfahren noch einmal definiert.

Über das Symposiumsprogramm hinaus wird hier auch ein Beitrag von R. Hofer und K.-D. Hänsgen zur Frage der **Trainierbarkeit von Testleistungen** aufgenommen. Voraussichtlich würde im Falle einer Einführung des Tests auch in der Schweiz diskutiert werden, ob die Trainingsangebote verschiedener privater Anbieter die Chancengleichheit beeinträchtigen. Es erscheint daher dringlich, mit den Ergebnissen der vorliegenden Untersuchung den Versprechen solcher Anbieter betreffend Leistungsverbesserungen eine kritische und sehr ernüchternde Sicht entgegenzustellen.

Es ist die Meinung der Veranstalter und Herausgeber dieses Bandes, dass ein Eignungstest unter Berücksichtigung der schweizerischen Verhältnisse letztlich die Methode der Wahl für die Zulassung zum Medizinstudium darstellt. Der hier zu verwendende Test ist wissenschaftlich sehr gut abgesichert und liegt in drei äquivalenten Sprachformen vor. Chancengleichheit ist vor allem durch eine standardisierte Durchführung und Auswertung voll gegeben. Mögen die hier vorgetragenen Argumente und Fakten die Diskussion in der Schweiz weiter voran bringen.

K.-D. Hänsgen N. Ischi

Fribourg/Bern, Dezember 1996

Eignungskriterien bei der Zulassung zum Medizinstudium in Europa: Ergebnisse einer Erhebung in allen europäischen Ländern

G. Trost

1. Ziele und Anlage der Erhebung

Im Jahr 1995 führten Judith Ebach und ich in allen europäischen Ländern eine Erhebung über die Zulassung zum Medizinstudium durch. Die Association of Medical Schools of Europe (AMSE) hatte die Anregung zu dieser Untersuchung gegeben, und eine Arbeitsgruppe dieser Vereinigung unter der Leitung von Professor Sergio Curtoni, Turin, unterstützte uns bei der Gewinnung kompetenter Ansprechpartner (in der Regel Dekane der medizinischen Fakultäten oder Vertreter nationaler Vereinigungen von Hochschullehrern der Medizin in den einzelnen Ländern). Gefördert wurde die Untersuchung durch die Europäische Kommission im Rahmen des ERASMUS-Programms.

Zwei Ziele wurden mit der Erhebung verfolgt:

a) Die Verfahrensweisen, die Auswahlkriterien und die diagnostischen Instrumente, welche bei der Zulassung zu den medizinischen Hochschulen herangezogen werden, sollten für jedes einzelne Land ermittelt und in übersichtlicher Form dargestellt werden.

b) Die Erfahrungen mit den verschiedenen Zulassungskriterien und Auswahlinstrumenten in den einzelnen Ländern sowie die vorliegenden Ergebnisse empirischer Evaluationsstudien sollten gesammelt und in einer länderübergreifenden Analyse ausgewertet werden.

Dank einer Reihe von Mahn- und Nachfaßaktionen lagen schließlich Informationen aus 35 europäischen Ländern, in denen es medizinische Hochschulen gibt, vor.

In diesem Beitrag wird ein Überblick gegeben über die Kriterien, die in den europäischen Ländern bei der Entscheidung über die Zulassung zum Medizinstudium angelegt werden. Ferner werden die Ergebnisse empirischer Studien zur Evaluation der wichtigsten Auswahlinstrumente zusammenfassend dargestellt. Die vollständige Version der Studie erscheint in diesen Wochen als Buch (Ebach & Trost, 1996).

2. Drei Grundmodelle der Hochschulzulassung

Zunächst sollen drei Grundmodelle vorgestellt werden, denen sich die nationalen Zulassungsregelungen zuordnen lassen.

Beim Grundmodell I wird jedem Bewerber, der die formale Zulassungsberechtigung besitzt - das ist in der Regel der erfolgreiche Abschluss der Sekundarstufe -, der unmittelbare Zugang zum Hochschulstudium eröffnet. Übertrifft die Zahl der Bewerber die Zahl der Plätze, die in den betreffenden Studiengängen auf Dauer angeboten werden können, so erfolgt die Auswahl während des Studiums, meist nach dem ersten, seltener nach dem zweiten Studienjahr, anhand von Hochschulprüfungen. Die Auswahl kann einstufig oder mehrstufig erfolgen.

Zu den Repräsentanten dieses Grundmodells zählen Frankreich, Österreich und, zumindest noch in diesem Jahr, die Schweiz. Auch Belgien hing bisher diesem Modell an; dort wird allerdings vom kommenden Jahr an im flämischen Teil vor Studienbeginn ein Testverfahren zur Auswahl unter den Medizinbewerbern durchgeführt werden (vgl. den Beitrag von Piet Janssen in diesem Bericht).

Bei den Grundmodellen II und III erfolgt die Auswahl vor der Zulassung.

Grundmodell II beruht auf der Annahme, dass sämtliche Bewerber, welche die schulische Sekundarstufe erfolgreich durchlaufen haben, grundsätzlich für ein Hochschulstudium geeignet seien. Im Falle eines Bewerberüberhangs wird deshalb eine Zufallsauswahl durch das Los getroffen. In reiner Form ist dieses Prinzip allerdings nirgendwo verwirklicht. In der Bundesrepublik Deutschland war ein „leistungsgesteuertes Losverfahren", bei dem die Loschancen von der Höhe der Abiturdurchschnittsnote mitbestimmt waren, Bestandteil des Übergangsverfahrens für die Zulassung zu den medizinischen Studiengängen in den Jahren 1980 bis 1985. In den Niederlanden findet in den sogenannten „Numerus-fixus-Fächern" ein gestuftes Verfahren statt, bei dem die Auswahl sowohl nach Leistung als auch nach Zufall erfolgt: Die erste Stufe bildet die Auswahl nach dem Notendurchschnitt in der Schule, in der zweiten Stufe wird ein Losverfahren angewendet.

Weitaus am häufigsten, nämlich in 31 der 35 Länder, wird bei der Zulassung zum Medizinstudium in Europa - wie auch ausserhalb Europas - nach dem Grundmodell III verfahren: Hier erfolgt die Zulassung nach den Kriterien der Eignung und Leistung; diese werden vor Studienbeginn auf verschiedene Weise überprüft. Innerhalb dieses Modells lassen sich zentrale und dezentrale Verfahrensweisen unterscheiden. Im ersten Falle werden die Zulassungskriterien in einheitlicher Weise festgelegt; sie sind somit für alle Hochschulen verbindlich, beispielsweise in den Numerus-clausus-Studiengängen in der Bundesrepublik Deutschland. Im zweiten Falle legt die einzelne Hochschule die Krite-

rien für die Zulassung fest und entscheidet in autonomer Weise; ein Beispiel für diese Praxis ist Grossbritannien.

In einem ersten Resümee lässt sich mithin festhalten: Praktisch überall in Europa findet eine Auswahl unter den Bewerbern für medizinische Studiengänge statt. In den meisten Ländern wird die Auswahlentscheidung vor der Zulassung zum Studium getroffen.

Im folgenden Abschnitt werden diejenigen europäischen Länder betrachtet, in denen eine Auswahl vor Studienbeginn stattfindet. Das Interesse richtet sich hierbei auf die Frage, anhand welcher Kriterien ausgewählt wird.

3. Auswahlkriterien bei der Zulassung zu den medizinischen Hochschulen in Europa

Fragt man zunächst nach der *Einheitlichkeit* der Zulassungsverfahren innerhalb der einzelnen europäischen Ländern, so stellt man fest, dass in 12 Ländern jeweils an allen medizinischen Hochschulen die gleichen Auswahlkriterien gelten; in weiteren neun Ländern sind die Kriterien jeweils an den meisten medizinischen Hochschulen ähnlich, in drei Ländern (Bulgarien, Schweden und Tschechien) sind sie jeweils unterschiedlich. Sechs weitere Länder haben jeweils nur eine medizinische Hochschule. (Für Rumänien fehlt die betreffende Information.)

Häufigstes Auswahlkriterium ist der *Leistungsnachweis im Abschlusszeugnis der Sekundarstufe*: In 23 europäischen Ländern wird dieser Indikator der Studieneignung bei der Zulassungsentscheidung herangezogen. In 12 Ländern wird die Belegung bestimmter - meist naturwissenschaftlicher - Fächer in der Sekundarstufe gefordert (Tabelle 1).

Das zweithäufigste Auswahlkriterium ist das Abschneiden in schulstoffbezogenen *Kenntnistests*.

In 22 europäischen Ländern spielen derartige Testergebnisse eine Rolle bei der Zulassung zum Medizinstudium. Meist sind die Inhalte der Schulfächer Biologie, Chemie und Physik Gegenstand der Prüfung durch die Tests; etwa in jedem vierten dieser Länder werden zudem die Ergebnisse von Tests in Mathematik oder Statistik, im Umgang mit der Muttersprache oder in einem Fremdsprachentest berücksichtigt. Ausser reinem Fachwissen wird teilweise auch das Verständnis für die jeweiligen Themengebiete geprüft.

In 18 Ländern bestehen diese Kenntnistests zumindest teilweise aus Aufgaben mit Mehrfachwahl-Antworten (Multiple-Choice-Items). Die Dauer der Tests

variiert beträchtlich; sie reicht von einer Stunde (an manchen italienischen Universitäten) bis zu 19 Stunden (in Spanien).

Studierfähigkeitstests zielen im Unterschied zu schulstoffbezogenen Kenntnistests nicht auf die Messung von in der Schule erworbenem Wissen, sondern auf die Erfassung kognitiver Fähigkeiten ab, die für die Bewältigung der Anforderungen des betreffenden Studiengangs besonders wichtig sind. Solche Testverfahren werden in fünf europäischen Ländern verwendet. In der Bundesrepublik Deutschland müssen sich derzeit alle Bewerber um medizinische Studienplätze einem studienfeldspezifischen Studierfähigkeitstest, dem Test für medizinische Studiengänge, unterziehen. In der Türkei bildet ein allgemeiner Studierfähigkeitstest die erste Stufe in einem zweistufigen Testprogramm; die zweite Stufe besteht aus fachbezogenen Kenntnistests. In Schweden können die Bewerber wählen, ob sie sich der Auswahl aufgrund des Ergebnisses in einem Studierfähigkeitstest oder aufgrund ihres Schulabgangszeugnisses stellen (vgl. den Beitrag von Ingemar Wedman in diesem Bericht). In Finnland und in Tschechien verwenden einzelne Hochschulen die Ergebnisse von Studierfähigkeitstests als Zulassungskriterien.

Mit Ausnahme des finnischen Tests bestehen die in Europa verwendeten Studierfähigkeitstests ausschliesslich aus Aufgaben mit Mehrfachwahl-Antworten. Die Dauer ist unterschiedlich. Der schwedische Test beispielsweise beansprucht etwas mehr als vier Stunden, der deutsche Test fünf Stunden.

Die Ergebnisse von *Interviews* werden in sieben europäischen Ländern, wenn auch nicht immer an allen medizinischen Hochschulen, bei der Auswahlentscheidung herangezogen (Dänemark, Deutschland, Grossbritannien, Schweden, Tschechien, Ukraine, Weissrussland). Die Zahl der beteiligten Personen unterscheidet sich von Land zu Land und teilweise von Universität zu Universität: In Grossbritannien stehen einem Bewerber bis zu fünf Interviewer gegenüber, in der Bundesrepublik Deutschland trifft ein Kandidat stets auf zwei Gesprächspartner, am Karolinska-Institut in Stockholm finden zwei Einzelgespräche mit je einem Interviewer statt.

Das Absolvieren eines *Krankenhauspraktikums* wird in keinem europäischen Land verbindlich als Zulassungskriterium gefordert. In Dänemark und Weissrussland können Kandidaten durch den Nachweis derartiger praktischer Erfahrung jedoch ihre Zulassungschancen verbessern. In Norwegen gilt dies für jegliche Art von Berufserfahrung. In der Bundesrepublik Deutschland können Bewerber durch das Absolvieren einer Berufsausbildung sowie durch das Ableisten eines mindestens einjährigen sozialen Dienstes während der „Wartezeit" ihre Position in der „Warteschlange" verbessern.

Die folgenden Kriterien werden jeweils nur in einem europäischen Land berücksichtigt: Die Wartezeit (Deutschland), ein *Empfehlungsschreiben* durch die

Schulleitung bzw. ein *„Statement of Interest"*, abgegeben durch die Bewerber selbst (Grossbritannien), der Erfolg in einem *naturwissenschaftlichen Wettbewerb* (Kroatien), das Ergebnis eines *Losverfahrens* (Niederlande), ein Zertifikat der *„Fähigkeit zur interpersonellen Kommunikation"* (Portugal).

Unter allen europäischen Ländern, in denen Zulassungsbeschränkungen herrschen, gibt es nur eines (Slowenien), in dem die Auswahl aufgrund eines einzigen Kriteriums erfolgt: den Noten in der Sekundarstufe. Diese Regelung gilt seit 1995; zuvor wurden ausserdem die Ergebnisse in Kenntnistests berücksichtigt. In allen anderen Ländern werden die Studierenden jeweils anhand einer *Kombination mehrerer Kriterien* ausgewählt.

Als Beispiel für ein besonders differenziertes Zulassungssystem, bei dem in einem Quotenmodell vier verschiedene Kriterien berücksichtigt werden, sei das Auswahlverfahren in der Bundesrepublik Deutschland für die medizinischen Studiengänge im Überblick dargestellt [1] :

– 45 Prozent der Plätze werden für Bewerber mit den höchsten Punktwerten vergeben, die sich aus der Kombination von Abiturdurchschnittsnote und Ergebnis im Test für medizinische Studiengänge im Verhältnis 55 zu 45 errechnen;

– 10 Prozent der Plätze sind Bewerbern mit den besten Testergebnissen ohne Rücksicht auf ihre Abiturdurchschnittsnote reserviert;

– 20 Prozent der Plätze werden an Bewerber vergeben, welche die längste Wartezeit aufzuweisen haben;

– 15 Prozent der Plätze werden aufgrund der Ergebnisse von Auswahlgesprächen, die durch die Universitäten durchgeführt werden, vergeben;

– 10 Prozent der Plätze sind für bestimmte Bewerbergruppen wie z.B. Zweitstudienbewerber sowie für „Härtefälle" vorgesehen (Trost, 1989).

Die mit Abstand *häufigste Kombination* von Zulassungskriterien ist jene von Schulnoten und *schulstoffbezogenen Kenntnistests*: In 15 europäischen Ländern ist diese Kombination zu finden; in sechs dieser Länder werden zusätzlich weitere Kriterien herangezogen.

[1] In der Bundesrepublik Deutschland wird das hier beschriebene Auswahlverfahren für die medizinischen Studiengänge zum Sommersemester 1998 durch ein vereinfachtes Verfahren abgelöst werden. Zwar haben die Bundesländer die Fortführung von Zulassungsbeschränkungen beschlossen - vor dem Hintergrund der drastischen Sparbeschlüsse spach sich jedoch eine Mehrheit der Länder für das weniger aufwendige Auwahlverfahren anhand der Kriterien Abiturdurchschnittsnote und Wartezeit aus.

4. Evaluation der Auswahlinstrumente

Eine Analyse der Verfahrensweisen bei der Hochschulzulassung in den europäischen Ländern bliebe unbefriedigend, liesse sie es bei der blossen Beschreibung des derzeit Praktizierten bewenden. Neben Antworten auf die Frage: „Was geschieht in den einzelnen Ländern?" sind Antworten auf eine zweite Frage von Interesse: „Wie gut funktioniert das, was in den einzelnen Ländern praktiziert wird?" Deshalb baten wir bei unserer Erhebung die Informanten in den einzelnen Ländern auch um eine persönliche Einschätzung des jeweils geltenden Zulassungsverfahrens sowie um die Überlassung aller Arten von Studien zur Evaluation der einzelnen Verfahren. Daneben zogen wir die in der internationalen Fachliteratur zugänglichen Ergebnisse von Bewährungskontrollen im Zusammenhang mit der Hochschulzulassung heran. Die Ergebnisse seien in diesem Abschnitt in knapper Form dargestellt.

Eine Hilfe zur Bewertung der einzelnen Auswahlinstrumente geben die klassischen Gütekriterien, die in der pädagogischen Diagnostik verwendet werden:

Die *Objektivität* eines Auswahlinstruments bezeichnet das Mass, in dem alle Bewerber in bezug auf die Durchführung und Auswertung eines Verfahrens sowie die Interpretation ihrer Reaktionen in diesem Auswahlverfahren gleich behandelt werden.

Die *Zuverlässigkeit* eines Auswahlverfahrens ist das Mass an Genauigkeit, mit dem die festzustellenden Eigenschaften erfasst werden; ein zuverlässiges Verfahren führt beispielsweise zu identischen Ergebnissen, wenn die Messung wiederholt wird.

Die *Gültigkeit* eines Auswahlinstruments beschreibt das Mass, in dem ein Verfahren in der Tat das misst, was es messen soll. Im vorliegenden Zusammenhang ist die prognostische Gültigkeit von besonderer Bedeutung; sie gibt das Ausmass an, in dem anhand der Ergebnisse des Auswahlverfahrens die spätere Studienleistung vorhergesagt werden kann. Bei der Bewertung der prognostischen Gültigkeit eines Auswahlinstruments muss jedoch stets auch die Güte der verfügbaren Erfolgskriterien - hier also der Indikatoren des Studienerfolgs - berücksichtigt werden: Nur wenn diese Kriterien ihrerseits hinreichende inhaltliche Gültigkeit und Zuverlässigkeit aufweisen, ist eine befriedigende Prognose aufgrund irgendeines Auswahlinstruments möglich.

Nicht nur an diesen wissenschaftlichen sondern auch an einigen weiteren Kriterien lässt sich die Brauchbarkeit eines Auswahlverfahrens messen. Dazu gehören:

- die *Ökonomie* des Auswahlverfahrens, also die Frage nach dem Verhältnis von Aufwand für die Entwicklung, Durchführung und Auswertung des Ver-

fahrens und dem Nutzen, der beispielsweise daraus erwächst, dass mit Hilfe dieses Verfahrens Fehlentscheidungen bei der Auswahl verringert werden.

- die *Fairness* des Auswahlverfahrens, das heisst die Vermeidung einer systematischen Benachteiligung bestimmter Bewerbergruppen, die im Mittel gleich geeignet sind wie andere Gruppen, durch die Verwendung der Auswahlkriteren.

- die *Akzeptanz* eines Auswahlverfahrens, also das Mass, in dem dieses Verfahren die Zustimmung von Studienbewerbern, Hochschulen, von Politik und Verwaltung im Bildungsbereich sowie in der Öffentlichkeit insgesamt findet.

- Schließlich sollte ein Auswahlverfahren stets auch darauf überprüft werden, ob es unerwünschte *Auswirkungen etwa auf das Bewerberverhalten* oder auf die Lehre in der Sekundarstufe hat.

Nicht zu allen dieser Kriterien liegen für die einzelnen Auswahlverfahren empirische Befunde vor. Die Bewertung der Auswahlverfahren wird sich deshalb im folgenden auf die Kriterien der Objektivität und Zuverlässigkeit, der prognostischen Gültigkeit, der Ökonomie und der Akzeptanz beschränken, und sie wird sich, der Übersichtlichkeit halber, auf die am häufigsten verwendeten Auswahlelemente: Schulnoten, Kenntnis- und Studierfähigkeitstests sowie Interviews beziehen.

4.1. Schulnoten

Schulnoten gelten primär als ein Mass des erreichten schulischen Leistungsstands, indirekt aber auch als Mass der intellektuellen Leistungsfähigkeit. Werden Schulnoten als Auswahlkriterien etwa bei der Hochschulzulassung verwendet, so ist zu bedenken, dass diese auch andere als diagnostische Funktionen - insbesondere pädagogische - haben.

Schulnoten weisen lediglich eine mässige Objektivität im Sinne der Beurteilerübereinstimmung und eine recht niedrige Wiederholungszuverlässigkeit auf (Baron-Boldt, 1989; Ingenkamp, 1989). Es gibt zahlreiche Belege für die unterschiedliche Bewertung der Schulleistungen von Lehrer zu Lehrer, von Schule zu Schule und von Land zu Land.

Trotz aller Schwächen des Beurteilungsverfahrens ist jedoch die Gesamtleistung in den letzten Schuljahren als vergleichsweise bestes Einzelmass zur Vorhersage des Studienerfolgs zu bezeichnen (Baron-Boldt, 1989; Trost, Klieme & Nauels, 1996).

Die Prognosekraft einzelner Fachnoten ist generell niedriger als diejenige des Notendurchschnitts im Abschlusszeugnis. Die einzelnen Fachnoten unterschei-

den sich in ihrer Prognosekraft hinsichtlich des Studienerfolgs zwar erheblich, aber selbst die höchsten Kennwerte für die prognostische Gültigkeit von Einzelnoten liegen deutlich unter denjenigen für die Durchschnittsnoten (Baron-Boldt, 1989; Weingardt, 1989).

Unter Kostengesichtspunkten stellen Schulnoten wegen ihrer leichten Verfügbarkeit ein sehr günstiges Auswahlkriterium dar.

Die Akzeptanz von Schulnoten als Auswahlkriterium bei der Hochschulzulassung ist mässig. Etwa die Hälfte aller Medizinbewerber in der Bundesrepublik Deutschland äusserten Anfang der achtziger Jahre eine positive Einstellung gegenüber diesem Auswahlkriterium, zugleich herrschte jedoch grosse Skepsis gegenüber der alleinigen Verwendung dieses Kriteriums (Trost, 1985).

4.2. Kenntnistests

Die Objektivität standardisierter schulstoffbezogener Kenntnistests, die im Rahmen der Hochschulzulassung eingesetzt werden, ist zumeist hoch. Das betrifft, dank einheitlicher Bedingungen der Testabnahme, sowohl die Objektivität der Durchführung als auch, soweit Aufgabenlösungen in Multiple-Choice-Form vorgegeben werden, die Objektivität der Auswertung. Weniger hoch ist die Objektivität von Kenntnistests mit offenen Antwortformen. Gleiches gilt für die Zuverlässigkeit.

Die Vorhersagekraft von Kenntnistests bleibt zwar, nach übereinstimmenden Ergebnissen der meisten Bewährungskontrollen, hinter derjenigen von Schulleistungen der Sekundarstufe zurück, ist aber immer noch zufriedenstellend (Alciati, Bonetto & Curtoni, 1993). Die Prognosekraft eines Kenntnistests steigt in der Regel mit der Ähnlichkeit von Test- und späteren Studieninhalten.

Durch die Kombination von Schulabschlussnoten und Ergebnissen in Kenntnistests kann die Vorhersage des Studienerfolgs verbessert werden gegenüber einer Vorhersage aufgrund eines der beiden Kriterien alleine.

Aufwand und Kosten für die Entwicklung und Erprobung standardisierter Kenntnistests sind relativ hoch. Sind die Tests jedoch einmal entwickelt und standardisiert, ist ihre Durchführung, auch mit grossen Teilnehmerzahlen, und ihre maschinelle Auswertung recht ökonomisch und viel weniger aufwendig als etwa die Durchführung von Interviews.

Generelle Aussagen über die Akzeptanz von Kenntnistests lassen sich aus länderübergreifenden Perspektive kaum treffen.

Kenntnistests können unerwünschte Auswirkungen auf Inhalte und Gestaltung des Schulunterrichts ausüben. In manchen Ländern wird beispielsweise die Qualität einer Schule anhand der Testergebnisse der Studienbewerber beurteilt,

weshalb die Schulen bemüht sind, ihre Schüler möglichst gut auf diese Tests vorzubereiten. Wenn die Unterrichtsinhalte in erster Linie an den Inhalten der Tests orientiert sind und der Schwerpunkt auf rein reproduktive Fähigkeiten gelegt wird, geht dies zu Lasten des Erwerbs eines breiteren Wissens und tieferen Verständnisses. Ferner können Kenntnistests das Aufblühen einer kommerziellen „Bildungsindustrie" nach sich ziehen, wenn der Schulunterricht den dort geforderten Prüfungsstoff nicht oder nicht ausreichend vermitteln kann. Schließlich ist auf die vergleichsweise hohe kurzfristige Trainierbarkeit der Leistungen in Kenntnistests, z.B. auch in kommerziellen Vorbereitungskursen, hinzuweisen.

4.3. Studierfähigkeitstests

Bei Studierfähigkeitstests mit Multiple-Choice-Antworten wie dem schwedischen oder dem deutschen Hochschuleingangstest ist eine sehr hohe Objektivität der Testdurchführung sowie der Testauswertung gegeben. Aber auch beim finnischen Studierfähigkeitstest, der ausschliesslich Fragen mit freien Antworten enthält, konnte dank der Standardisierung der Auswertungsschemata und des Trainings der Auswerte eine 95prozentige Übereinstimmung der Auswerte erzielt werden (Lindblom-Ylänne & Lonka, 1995).

Hohe Werte, die den international üblichen Anforderungen entsprechen, haben sich auch für die Zuverlässigkeit der Ergebnisse des deutschen Tests für medizinische Studiengänge nachweisen lassen (Deidesheimer Kreis, 1996; Trost, 1995a).

Die prognostische Gültigkeit von Studierfähigkeitstests ist etwas niedriger als jene der Schulnoten, aber annähernd gleich hoch wie die Gültigkeit von Kenntnistests (Beller, 1993; McDonald, 1975; Trost, 1995b; Whitney, 1989).

Wenn Schulnoten in der Sekundarstufe kombiniert werden mit den Ergebnissen in Studierfähigkeitstests, ist der Zuwachs an prognostischer Gültigkeit in den meisten Fällen höher, als wenn Schulnoten mit den Ergebnissen in schulstoffbezogenen Kenntnistests kombiniert werden.

Was die Ökonomie der Verwendung von Studierfähigkeitstests betrifft, ist die Situation ähnlich derer für Kenntnistests. Die Kosten für die Entwicklung und Erprobung standardisierter Testverfahren sind hoch; die Durchführung mit grossen Teilnehmerzahlen ist vergleichsweise kostengünstig, und die maschinelle Auswertung ist billig. Bei den Überlegungen zum Verhältnis von Aufwand und Nutzen ist jedoch stets auch der mögliche Gewinn an Treffsicherheit bei der Auswahl von Studierenden und damit die Verminderung der Zahl von Fehlbelegungen, Studienabbrechern oder Langzeitstudierenden zu bedenken.

Erfahrungen über die Akzeptanz liegen für den deutschen Test für medizinische Studiengänge vor. Etwas mehr als die Hälfte der Medizinbewerber betrachten die Verwendung dieses Tests als Verbesserung des Zulassungsverfahrens (Trost, 1993a).

4.4. Interviews

Die Objektivität des Interviews als Auswahlinstrument kann bestimmt werden als das Mass der Übereinstimmung der Urteile von zwei oder mehr Interviewern über dieselben Kandidaten. Aus den europäischen Ländern liegen für das Interview als Instrument der Zulassung zum Medizinstudium keine diesbezüglichen Informationen vor. In einem Überblick über amerikanische Untersuchungen fanden Edwards, Johnson und Molidor (1990) Kennwerte für die Beurteilerübereinstimmung, die über eine weite Spanne streuten, aber überwiegend im niedrigen Bereich lagen. Die allgemeine Interviewforschung zeigt, dass die Objektivität des Interviews beträchtlich erhöht werden kann, wenn den Interviewern detaillierte Leitfäden an die Hand gegeben werden, wenn die zu ermittelnden Eignungsmerkmale klar definiert sind, wenn klare Regeln für die Auswertung der Gespräche vorgegeben werden und, nicht zuletzt, wenn die Interviewer für diese Tätigkeit geschult werden (Trost, 1996).

Die prognostische Gültigkeit von Interview-Ergebnissen ist viel niedriger als diejenige der bisher diskutierten Auswahlkriterien (Edward, Johnson & Molidor, 1990; Trost, 1996). Dies ist teilweise die Folge der überwiegend unzureichenden Objektivität der Interviews.

Verwendet man das Interview zusätzlich zu anderen diagnostischen Verfahren, ist der Gewinn an prognostischer Gültigkeit ebenfalls gering.

Es gibt jedoch Hinweise darauf, dass der Erfolg im klinischen Teil der medizinischen Ausbildung etwas besser auf der Basis von Interviews vorhergesagt werden kann als der Erfolg im vorklinischen Studium (z.B. Hall, Regan-Smith & Tivnan, 1992).

Auch bezüglich der Prognosekraft des Interviews sind Verbesserungen durch die oben beschriebenen Massnahmen möglich, ferner durch die Beschränkung auf die Beurteilung derjenigen Merkmale, die im Interview vergleichsweise gut, unter Umständen sogar besser als mit anderen Prüfverfahren, erfaßbar sind. Dazu gehören beispielsweise Fähigkeiten zur sozialen Interaktion und zur Kommunikation, also Merkmale, die in der Gesprächssituation unmittelbar beobachtbar sind.

Der Aufwand, den das Interview vor allem in personeller Hinsicht erfordert, ist erheblich; es ist mithin unter verfahrenstechnischen wie organisatorischen Gesichtspunkten wenig ökonomisch. Deshalb wird das Interview dort, wo es beim

Hochschulzugang verwendet wird, häufig auf eine kleinere Zahl von Bewerbern beschränkt, die bereits anhand anderer Kriterien vorausgelesen sind.

Ungeachtet seiner überwiegend unbefriedigenden psychometrischen Qualitäten erfreut sich das Interview einer vergleichsweise hohen Akzeptanz zumindest bei den Bewerbern. 77 Prozent der deutschen Medizinbewerber betrachten die Einbeziehung eines Auswahlgesprächs als Verbesserung des Zulassungsverfahrens (Trost, 1993a).

5. Zusammenfassung

Der Überblick über die Praxis der Zulassung zum Medizinstudium in den Ländern Europas zeigt eine erhebliche Vielfalt der verwendeten Auswahlkriterien und Auswahlinstrumente. Zwölf Verschiedene Arten von Zulassungskriterien werden in unterschiedlichen Kombinationen herangezogen.

Die häufigste Kombination ist die Berücksichtigung der Schulabschlussnoten sowie der Ergebnisse eines schulstoffbezogenen Kenntnistests; in der Rangreihe nach Häufigkeit folgt die Kombination von Schulnoten mit der Forderung nach der Belegung bestimmter Schulfächer in der Sekundarstufe, dann die Kombination von Schulnoten und Interviews und die Kombination von Schulnoten und Studierfähigkeitstests.

Ein Streifzug durch die Literatur zur Evaluation der einzelnen oder der kombinierten Auswahlkriterien bzw. -instrumente zeigt: Betrachtet man diese Auswahlfaktoren im einzelnen, so kommt der Schulleistung in der Sekundarstufe die vergleichsweise höchste prognostische Gültigkeit in bezug auf den Studienerfolg zu. An zweiter Stelle rangieren Studierfähigkeitstests und schulstoffbezogene Kenntnistests. Kombiniert man Tests dieser Art mit schulischen Leistungsbeurteilungen, so lässt sich dadurch die Prognose des Studienerfolgs gegenüber der Verwendung eines jener Faktoren allein bedeutsam verbessern. Dem Interview, so wie es bei der Zulassung zumeist gehandhabt wird, kommt eine relativ geringe Vorhersagekraft bezüglich des Studienerfolgs zu. Es gibt jedoch Hinweise, wie sich sowohl die Objektivität als auch die prognostische Gültigkeit des Interviews verbessern lassen.

Literatur

Alciati, A., Bonetto, A. & Curtoni, S. (1993). La previsione del successo degli studenti in medicina et chirurgia. La Formazione del Medico, 8, 32-39.

Baron-Boldt, J. (1989). Die Validität von Schulabschlußnoten für die Prognose von Ausbildungs- und Studienerfolg. Frankfurt/M.: Peter Lang.

Beller, M. (1993). Zulassungsverfahren an israelischen Universitäten: Psychometrische und soziale Betrachtungen. In G. Trost, K. Ingenkamp & R.S. Jäger (Hrsg.), Tests und Trends 10. Jahrbuch der Pädagogischen Diagnostik (S. 115-142). Weinheim: Beltz.

Deidesheimer Kreis (1996). Hochschulzulassung und Studieneignungstests. Studienfeldbezogene Verfahren zur Feststellung der Eignung für Numerus-clausus- und andere Studiengänge. Göttingen: Vandenhoeck & Ruprecht.

Ebach, J. & Trost, G. (1996). Admission to Medical Schools in Europe. Overview on admission procedures - Evaluation of selection instruments - Samples of assessment elements. Lengerich/Berlin/Düsseldorf/Riga/Wien/Zagreb: Pabst Science Publishers.

Edwards, J.C., Johnson, M.S. & Molidor, J.B. (1990). The interview in the admission process. Academic Medicine, 65, 167-177.

Hall, F. R., Regan-Smith, M. & Tivnan, T. (1992). Relations of medical students' admission interview scores to their dean's letter ratings. Academic Medicine, 67, 842-845.

Ingenkamp, K. (1989). Diagnostik in der Schule. Beiträge zu Schlüsselfragen der Schülerbeurteilung. Weinheim: Beltz.

Lindblom-Ylänne, S. & Lonka, K. (1995). Selecting students for Medical School: What predicts success? A cognitive approach. University of Helsinky, Department of Psychology. Unpublished manuscript.

McDonell, W. (1975). Testing for student selection at tertiary level: A literature review. Victoria: Australian Council for Educational Research in association with Tertiary Education Entrance Project.

Trost, G. (1985). Pädagogische Diagnostik beim Hochschulzugang, dargestellt am Beispiel der Zulassung zu den medizinischen Studiengängen. In R. S. Jäger, R. Horn & K. Ingenkamp (Hrsg.), Tests und Trends 4. Jahrbuch der Pädagogischen Diagnostik (S. 41-81). Weinheim: Beltz.

Trost, G. (1989). A nationwide testing program for admission to medical schools in West Germany. In R.C. King & J.K. Collins (Eds.), Social applications and issues in psychology (pp. 131-137). Amsterdam: Elesevier Science.

Trost, G. (1993). Attitudes and reactions of West German students with respect to scholastic aptitude tests in selection and counselling programs. In B. Nevo & R. S. Jäger (Eds.), Educational and Psychological Testing: The Test Taker's Outlook (pp. 177-200). Göttingen: Hogrefe.

Trost, G. (Hrsg.). (1995a). Tests für medizinische Studiengänge (TMS): Studien zur Evaluation. 19. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Trost, G. (1995b). Principles and practice in selection for admission to higher education. In Kellaghan, T. (Ed.), Admission to Higher Education: Issues and Practice (pp. 7-15). Princeton, NJ: International Association for Educational Assessment.

Trost, G. (1996). Interview. In K. Pawlik (Hrsg.), Grundlagen und Methoden der Differentiellen Psychologie. Enzyklopädie der Psychologie (S. 464-505). Göttingen: Hogrefe.

Trost, G., Klieme, E. & Nauels, H.-U. (1996). The relationship between different criteria for admission to Medical School and student success. In D. Ajar (Ed.), New horizons in learning assessment. Ethical, strategic, methodological aspects. London, New York: Pergamon Press (in print).

Weingardt, E. (1989). Untersuchungen über Korrelationen zwischen Reifeprüfungsnoten und Erfolg auf der Universität. In K. Ingenkamp (Hrsg.), Die Fragwürdigkeit der Zensurengebung (S. 252-255). Weinheim: Beltz.

Whitney, D. R. (1989). Educational admissions and placement. In R.L. Linn (Ed.), Educational measurement (pp. 515-525). New York: American Council on Education & Macmillan Publishing Company.

(auf dieser Seite Tabelle 1 einfügen)

(auf dieser Seite Tabelle 2 einfügen)

The Swedish Scholastic Aptitude. Test Research and main findings

I. Wedman and W. Henriksson

1. Introduction

Today about 130,000 applicants for studies at the tertiary level sit the Swedish Scholastic Aptitude Test (SweSAT) each year. The test has almost become an "industry" in itself when it comes to selection to higher education. The regular staff responsible for the development of the test, or rather, the test battery, consists of 15 persons (in-house staff), but in total about 100 persons in different positions are involved in the development process. The test has become a regular part of the process of selection to higher education.

In this paper we will present the background of the test battery, some reflections and experiences obtained during the development process, basic issues that have been addressed and met during the existence of the test, and some remarks as to what should be considered when implementing a test like the SweSAT.

The SweSAT Programme

During the last three decade, the system of higher education in Sweden has been the subject of several commissions and committees as well as part of the general debate about education. Many changes have also been introduced during this period. Part of the interest in and the debate about the higher education system has dealt specifically with the selection problem, that is, how admission to higher education is organised and what rules and instruments characterise the selection system.

During the same period a Swedish admission testing program has been developed - the Swedish Scholastic Aptitude Testing Program, SweSAT for short. It has been used for selection purposes since 1977. Before that time, selection was solely based on the average applicant's mark from upper secondary school, provided that formal requirements were fulfilled for the study program in question. Such formal requirements were expressed as a minimum level based on marks or grades obtained in various subjects.

In 1977, the system was changed. For applicants who were older than 25 years and had more than four years of work experience, the SweSAT provided an opportunity to gain access to higher education in case there were more applicants than places and selection was necessary (see also Wedman, 1994).

Some background information

During the 1960s the higher education system began to change in Sweden. Since then several government commissions have discussed and proposed further changes of the system. A special aspect of interest has been the admission system, that is, the selection of applicants for many of the various study programs. The latest commission report was presented in 1985 and led to a major change of the system, including a great expansion of SweSAT use from 1991 onward.

A parallel and very important development has concerned the marking/grading system used in primary and secondary schools. From the early 1960s onwards the marking system in Sweden has was norm-referenced, that is grading on a curve. In 1995, this was changed to a criterion-referenced system. This change will eventually affect the selection system used when admitting students to universities and colleges. Given these circumstances, it is expected that the use of the SweSAT programme will expand further.

The SweSAT Programme

As part of the change in the higher education system in the 1960s, the admission system was reviewed. An expert group led by Professor Sten Henrysson began investigating the possibility of developing an entrance test to be used for all applicants to higher education. Many of the various tests applied in other countries were studied and put to trial use.

There were several arguments put forward for developing an entrance examination of the type that the SweSAT represents. One of the main arguments was the need for means by which to compare applicants who lacked comparable marks. Many of those applicants belonged to the so-called 25:4 group, that is, applicants who were at least 25 years old and had at least four years of work experience. There was a politically strong and well-articulated ambition to make colleges and universities accessible to more people.

Another argument for developing an entrance test was the idea that students from upper secondary school should be given a second chance to gain access to colleges and universities. At the time, only marks were used when selecting students for entrance to colleges and universities. A second route was sought to reduce the emphasis placed on marks or grades.

In 1975, the Swedish Parliament decided to change the admission system beginning with the autumn 1977 student selection. However, the decision provided that the use of the entrance test should be limited to applicants who were at least 25 years old and had at least four years of work experience. The argument regarding a second chance for younger students was to be reconsidered when the new system had been in operation for some time.

The decision to limit the use of the SweSAT program led to about 10,000 applicants taking the test each year from 1977 to 1989, about 6,000 at the spring administration of the SweSAT and 4,000 at the autumn testing.

As a result of the government commission report about changing the admission system was presented in 1985, the use of the SweSAT programme has increased dramatically. The commission suggested that the SweSAT programme should be open to all applicants, including those leaving upper secondary school. Beginning with the 1991 autumn administration of the SweSAT, selection to higher education is based either on marks from upper secondary school or on SweSAT results. So far, about per cent of the available places in study programs with a limited number of places have had applicants selected on the basis of average marks and about 40 per cent on the basis of SweSAT result.

Applicants are judged on the most favourable condition. They do not have to decide themselves whether to compete on the basis of their average mark or their SweSAT result, which has also meant that most applicants choose to take the test. About 130,000 students do so each year.

2. The present content of the SweSAT

The SweSAT has been used in selection to higher education since 1977. At first the SweSAT was open only to applicants who were at least 25 years old and who had at least four years of work experience. As from 1991 the SweSAT may be used by all applicants to higher education. The selection procedure is based on three groups of applicants; those with GPA (grade point average) from upper secondary school, those with SweSAT-score, and those with SweSAT-scores with additional scores for work experience (WE). Applicants can belong to more than one group and, if so, the student is selected in the group in which he/she has the best rank. About 60 per cent of the places available at a certain study programme go to applicants selected on the basis of GPA and about 40 per cent to applicants selected on the basis of test score.

The content of the SweSAT remained almost the same until 1996. Until then, the SweSAT consisted of six subtests and the items in a certain version (two versions are administered each year) were public documents as soon as the sixth and last subtest of a SweSAT administration had been completed by the test taker. Some minor changes of the SweSAT programme were conducted between 1977 and 1996. Worth mentioning is that in 1992, ERC, a reading comprehension test in English, was introduced. ERC replaced the subtest aiming at capturing the complex skills of study techniques.

Since 1996 the SweSAT consists of five subtests, as the general information subtest has been eliminated. The SweSAT is now administered in five blocks of

which one block consists of try-out items. The items in the try-out block are not public. The total test time is 50 minutes for each block, two blocks consist of the same subtests. The design of the administration is illustrated in Table 1 and the SweSAT programme is described in Table 2.

Table 1: Regular test items and try-out items in the SweSAT programme divided into blocks.

Block	**Subtest**	**Abbreviation**	**Time**
1	Interpretation of diagrams, tables and maps	DTM	50
2	English reading comprehension	ELF	35+
	Vocabulary	WORD	15
3	Reading comprehension	READ	50
4	Data sufficiency	DS	50
5	Any block of above		50

The order of the blocks (1-4) varies from one administration to another. The items in the fifth block are try-out items. These are not included in the test takers score.

Table 2: The Swedish Scholastic Aptitude Test (SweSAT).

Subtest	**Abbreviation**	**Number of items**	**Time**
Interpretation of diagrams, tables and maps	DTM	20	50
English reading comprehension	ELF	20	35
Vocabulary	WORD	40	15
Reading comprehension	READ	20	50
Data sufficiency	DS	22	50
Total		122	3 hrs 20 min

3. Five major problems encountered

During the development phase of the SweSAT Program we have encountered five major problem areas which may categorised as follows:

1. Social bias
2. Sex bias
3. Coaching
4. Prediction of academic performance
5. Practical matters

Each of these problem areas will be briefly described below.

3.1. Social bias

During the development phase, the issue of social bias in a test like the SweSAT Programme was heavily discussed, and primarily from a political point of view. It should be remembered that the test was introduced primarily to make higher education accessible to persons that earlier had been shut out from studies at the tertiary level. It is also important to note that Sweden this time had a social

democratic government. The very reason for introducing the SweSAT Programme was to find a way to improve opportunities for people that, generally speaking, represented lower socio-economic groups.

If one divides all applicants into socio-economic groups, all statistics show that those from the upper groups obtain a higher average score on a test like the SweSAT than those from lower socio-economic groups. The same picture will appear for almost every cognitive measure known of, including grades. The reason for this is quite simple and has primarily nothing to do with socio-economic status per se, but rather with the fact that applicants from higher socio-economic groups enjoy better study conditions than applicants from lower socio-economic groups.

To some influential Swedish politicians, this came as a bite of a surprise. The future of the test was up for discussion. However, there was, as it seemed, no third way out of the dilemma of introducing an admission test for older applicants (without comparable merits) and at the same time having a part of so called social bias in such a test. The discussion about social bias ended more or less and the idea of introducing the test gained new strength. When the test was introduced in 1977 it was supported by almost all interest groups.

In retrospect, social bias was not the issue of greatest concern when the test was introduced in 1977. Instead, the idea of measuring cognitive achievement by a test that took about four hours was. At this time, there was a very strong movement in Sweden for abolishing any written instrument that proposed to measure cognitive achievement. Students and teachers all over Sweden, primarily at the university level, questioned every form of written assessment and suggested instead group work and group assessment in its widest sense, a movement that ended in the beginning of the 80s.

The scientists that had developed the test were greatly concerned when it was introduced and feared massive criticism, given the circumstances mentioned above. However, nothing happened. No criticism, was voiced. The explanation for this has remained somewhat unclear, but perhaps the founder of the test, Professor emeritus Sten Henrysson, was right when he pointed to a number of significant events took place around the world on the very day that the test was introduced, and that these events caught most of the mass media interest.

3.2. Sex differences

In the 80's, the debate on social bias was replaced by the debate on gender. In fact the gender issue has influenced the work with the SweSAT Programme to a much greater extent than the social bias issue. The basic reason for this is that males generally speaking, outperform females on tests like the SweSAT. In the western part of the world, thousands of studies have been conducted in order to

find underlying causes. We have ourselves conducted a number of studies in this field. There seems to be no clear-cut explanation for this fact, even though a number of hypotheses have been presented.

The difference is rather small but systematic. It is important to bear in mind that the difference appears on a group level. In fact the distributions of males and females overlap to a very great extent, which means that a huge number of females outperform males on the SweSAT. Yet, on a group level, males outperform females.

One hypothesis presented is that the very format of the SweSAT and similar test batteries works in favour of males, i.e. males seem to handle the multiple-choice format better than females do. This is, however, a question under debate. On the other hand, we know that females usually outperform males on essays. Results from a number of studies indicate that this is the case, although the reasons remain unclear. Therefore, to incorporate an essay part in the SweSAT Program would likely lead to a test battery in which the differences in achievement between males and females would vanish. What makes the idea of incorporating an essay part unrealisable within the SweSAT Programme and most other programmes of this kind, are the costs involved. In order to obtain reliable results in scoring an essay you need at least three independent scorers, which in our case means 390,000 scorers per year!

The issue of gender differences should, according to our experiences, be approached with great seriousness. Besides investing a great amount of money in doing our own research in the field (Stage 1993), we have taken measures to make sure that men and women are equally represented in different expert panels that examine the test before it is presented. Since a few years back the government group for equal treatment (of males and females) is free to send a member to participate in the development and the final examination of a new test version.

3.3. Coaching

Henriksson (1981) made a review of relevant studies of the effects of practice and coaching on test score and his conclusion was that the effects caused by low test experience and a complex item format are the two main factors leading to score gains.

Two different approaches to the issue of special preparation and coaching. The first approach is to ensure that all test takers are familiar with the test and with the strategies required to take it. This approach will be discussed below under the heading "The SweSAT and Test-Wiseness".

The second approach is to develop tests free from items susceptible to special training, cramming, and coaching. This latter approach will be discussed in the next section in connection with Data Sufficiency items.

Data Sufficiency (DS) and Item Format

It has often been claimed that items with a complex format are susceptible to coaching (Vernon, 1960; Pike & Evans, 1972; Powers, 1986; Becker, 1990):

The results of this review support the hypothesis that complex item formats are indeed more susceptible to coaching, practice, and/or test preparation than are simpler formats. The number of words used to convey directions seems to be a significant aspect of this complexity, as does the fixed format nature of many of the most susceptible item types (Powers, 1986, p 76).

However, increasing an examinee's familiarity with novel item types (such as the SAT-M's datasufficiency or quantitative-comparison items) may well enable him or her to improve SAT performance considerably (Becker, 1990, p. 404).

In a series of studies Henriksson (1981) investigated DS items and their susceptibility to special instruction and coaching. The basic idea for the coaching intervention was that a rational problem-solving behaviour for DS items could be modelled as a sequential process including 2 (or 3) subproblems, i.e. 2 (or 3) decisions (Yes/No) in relation to each subproblem. The DS item format was also reduced in complexity by this transformation of the original problem to 2 (or 3) simpler subproblems which were to be answered separately. The coaching also included information about how to use partial knowledge.

A brief summary of the results of these three studies is that the test score was affected neither by coaching nor by practice. These conclusions also applied to conventional scoring (0-1) as well as to differential scoring. The purpose of the differential scoring model was to reflect partial knowledge.

In three studies (Henriksson, 1981) a special item format was developed with the instructed strategy as a main point of departure. This item format made the test taker act automatically according to the sequential process of problem-solving which was instructed orally in the earlier studies. The results of these three studies indicated that the obtained score was not affected by item format; neither if the score was based on conventional scoring nor on differential scoring.

To sum up, the conclusion drawn from the studies by Henriksson was that the sequential problem-solving strategy, which was instructed as well as induced in the item format in order to obtain a logical and rational problem-solving procedure, had no effect. Regarding the question of the complexity of the DS item format the results also indicated that the DS format is not complex, at least not

for students from grade three in upper secondary school. This means that the DS item format optimises the relation between actual knowledge and obtained score for a test taker, i.e. the score reflects the test taker's, capacity in terms of full as well as partial knowledge.

The SweSAT and Test-Wiseness

If test takers have little or no experience of tests and test taking, their obtained scores may be too low and not in accordance with their actual capacity. A fundamental assumption for test-fairness in this respect is that the test takers are test-wise. Consequently, the required test-wiseness should be made available to all potential test takers. With reference to the SweSAT, there are two different approaches to the problem of test-wiseness.

Everyone who registers for the SweSAT gets the information brochure "Högskoleprovet". The aim of this material is to familiarise the test takers with the SweSAT and to inform them about different circumstances and details before, during and after the testing. The material also presents the item format and sample items for each of the subtests, and the potential test taker is also requested to take this mini-SweSAT within the given time.

Another circumstance, contributing to test-wiseness and therefore worth mentioning, is that four out of five administered blocks of SweSAT-items are public documents. This means that all items, except items in the try-out block, are public. Consequently a potential test taker can get access to, for example, the latest SweSAT version and, thus, may practise on a full sample test.

The summarised judgement is that access to older versions of the SweSAT, together with the familiarisation material, create optimal conditions for test-fairness for the SweSAT with respect to test-wiseness.

3.4. Prediction of academic success

Attempts to predict academic success have gained much attention and several studies have been carried out to get information about the design of the admission procedure. Most universities utilise selection procedures based on individuals, previous performance, such as college grades (GPA) or results on tests, such as SAT, ACT or SweSAT, which are designed to predict general academic success.

Predictive studies are important for the evaluation of the strength of an admissions test like the SweSAT. As a consequence, thousands of predictive validity studies have been carried out over the years within the field of educational measurement, e.g. Willingham et al (1990); Henriksson & Wedman (1992). The majority of these studies have been reported as a single correlation coefficient

with the grades obtained after one or a few years of academic studies. However, the correlations and the results of predictive studies are often very difficult to interpret because of a number of statistical and other problems. Only admitting those who are at the top of the ability scale leads, for example, to the well-known restriction of range problem. The reliability of the criterion as well as a very restricted criterion scale, which is the case with using pass/fail grading scales, are other problems with predictive studies -(Henriksson & Wedman, 1992).

SweSAT and predictive studies

In order to get information about the predictive validity of the SweSAT, a rather elaborate predictive study was carried out in the seventies. This study was based on a predecessor to the SweSAT, e.g. Henrysson & Wedman (1979). The subjects (n=584) were students of engineering at two technological institutes and students at teacher training colleges. Test score, admission points (GPA) and a relatively refined criterion scale, based on a summary of a number of exams, were collected. Information was also collected on, among other things, study results, change of schools, and study interruptions. After the first year information was obtained about the points acquired on each subcourse. After the second year the same information was collected again. This concluded the data collection for the technical colleges, while information was once more - three years later - obtained from the teacher training colleges. The results obtained showed a fairly high correlation (about 0.4-0.5) between the test and the criterion, which was interpreted as a support for the hypothesis that SweSAT could compete with grades in predicting academic performance.

Henrysson et al (1985) carried out a validity study which consisted of 12 study programmes at four universities (n=3 171). The main result of the study was that students with three years of upper secondary school was the best performing group. Academic success was defined as the total number of credits achieved by the students. On the whole, GPA and SweSAT showed the expected correlation with academic success, i.e in the interval 0.35-0.45. Another result was that work experience (WE) seemed to reduce the predictive validity of the admission instrument (GPA or SweSAT-score). Three explanations were put forward. (1) Students admitted by way of WE often had a lower GPA than students without WE. (2) Students admitted by way of WE were older than other students and lacked previous experience of studies. (3) The older, WE-admitted students in general already had families and must provide for these. Such conditions could be thought to give WE-admitted students less time to concentrate on their studies.

In the early 90s new questions were raised about the predictive validity of the SweSAT, basically because its use in the selection procedure had been expanded

to allow all students to take the test. These changes created a need for a validity study of current interest. The results of validity studies are usually expressed as a correlation between the current predictor and some form of criterion of academic success. Since there are recurrent and difficult methodological problems associated with this type of design, i.e. Henriksson & Wedman (1992), a study was carried out with a different design. A sample of students from three different admission groups was followed during three years of academic studies. The study will be published in Scandinavian Journal of Education Research in 1997 (Henriksson & Wolming). In the following section we will highlight the main findings of this study.

In this study student performance during a period of three years was observed for a sample of 840 students from four different study programmes (Business administration, Medicine, Technical engineering physics, and Social works). All students at these study programmes at two out of five universities were selected. The students were divided into three groups, according to the grounds on which they had been admitted to the programmes. The first group of students had been admitted on the basis of earlier academic achievement (GPA), the second group on the basis of their SweSAT-scores, and the third group on the basis of their SweSAT-scores with additional scores for work experience (WE). The purpose of the study was to see if there were any differences in academic performance at the programmes between the three groups. Academic performance was defined in three ways:

a) the number of credits achieved after each semester

b) drop-out frequency after each semester

c) the number of credits achieved after each semester by those students who completed their studies.

Swedish universities and colleges use a system of credit points for studies on the undergraduate and graduate levels. One point corresponds to one week of full-time studies. One academic term comprises 20 weeks (20 points). 40 credits represent one full academic year. The number of credits acquired by all students was registered at the end of each semester. Information about study results, predictor in-formation (GPA, SweSAT, WE), and the grounds on which a certain student had been admitted to the programme was obtained from the national student records database.

* Credits achieved after each semester

The first criterion variable in the study was the number of credits acquired over time by the three admission groups. The drop-out frequency was not taken into account in this first analyse.

Figure 1 shows the performance of the three admission groups after each completed semester.

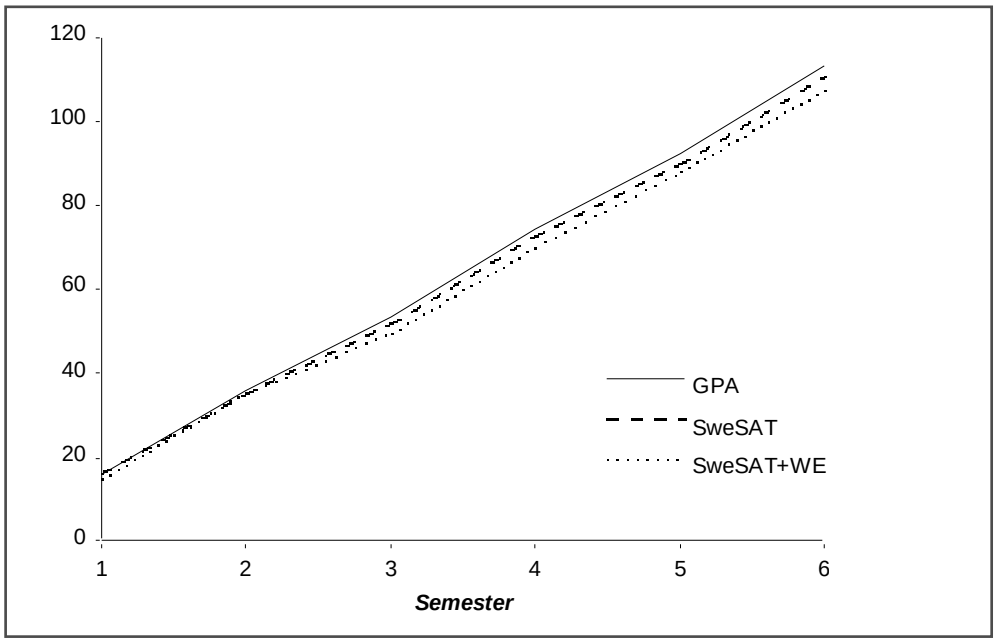


Figure 1. Number of credits acquired after each semester by the three admission groups. All study programmes.

At first glance, the differences between the admission groups as displayed in Figure 1 seem relatively small. However, the statistical analysis (Kruskal-Wallis 1-Way Anova) indicated a significant difference between the admission groups ($p<0.05$).

A pairwise follow-up analysis with the Mann-Whitneys U-test on all study programmes indicated that the differences were to be found between the students admitted on the basis of GPA and the students admitted on the basis of SweSAT+WE. The differences between the two groups appeared after semester 2, 3, 4, 5, and 6. Significant differences were also found between the two groups admitted on the basis of SweSAT (with and without WE). The differences between these two groups appeared after semester 2, 4, and 6. No significant differences were found between the students admitted on the basis of GPA and on the basis of SweSAT.

The analyses above were also made for each of the four study programmes. The significant differences which appeared between the GPA-group and SweSAT+-WE-group in the total group (all study programmes) could be related to the Business administration and economics study programme. At the other study programmes, no differences were found between the admission groups.

* Drop-out frequency after each semester

The second criterion variable was the drop-out frequency after each semester in the three admission groups. This is shown in Figure 2.

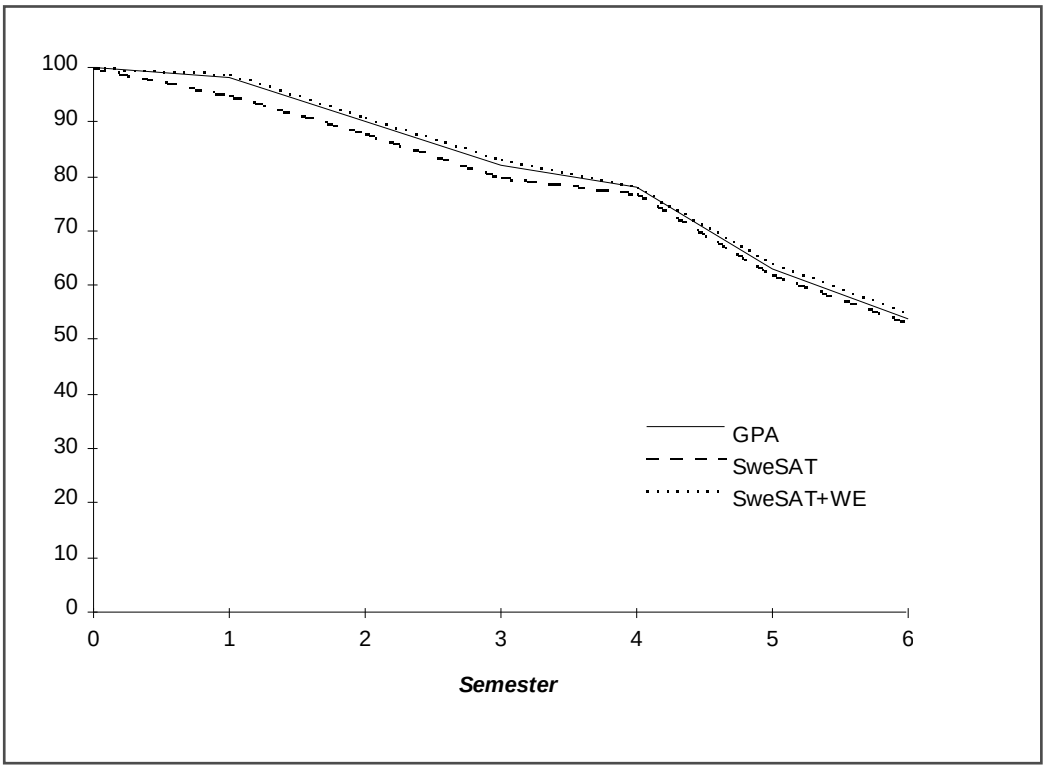


Figure 2. Drop-out frequency after each semester by the three admission groups. All study programmes.

As shown in Figure 2, the drop-out frequency is very much the same in the three groups. The Chi-Square analysis did not indicate any significant differences between the groups in this respect. When the study programmes were tested one by one, differences appeared at the Medical study programme. As shown in Figure 3, the drop-out frequency is higher in the WE-group than in the other two groups, and the dropping out takes place after the second semester. The figure also shows that dropping out takes place in the SweSAT-group after the third semester.

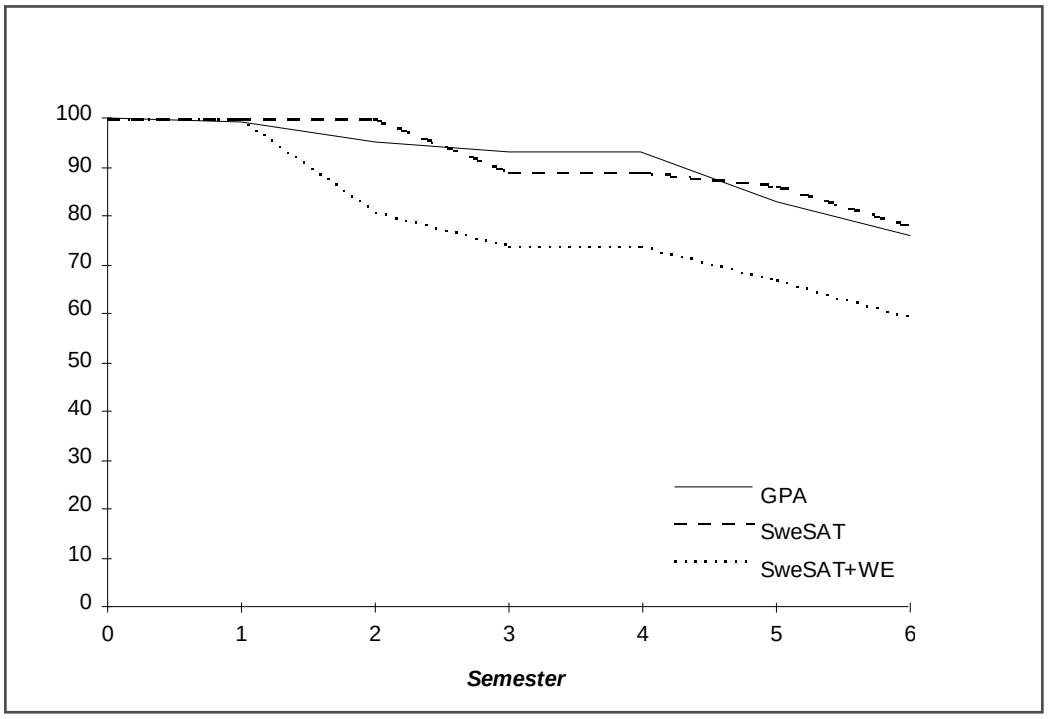


Figure 3. Drop-out frequency after each semester by the three admission groups. Medical study programme.

The drop-outs in the WE-group after the second semester consisted of five students. The drop-outs which were found in the SweSAT-group after the third semester consisted of three students. It can also be noted that after the second semester, the inclination of the drop-out curve was about the same for each of the three admission groups (Figure 3).

* Credits after each semester for those students who completed their studies

The third and final criterion variable used in this study was the number of credits acquired after each semester by the remaining students in the three admission groups. The main reason for making a distinction between success for the total group and success for the remaining group was that not considering drop-out frequency might give an incorrect picture of success if there are systematic differences between groups and study programmes.

The average number of credits acquired by the remaining students closely resembles the outcome of the first variable, which did not take the drop-out frequency within the groups into account (Figure 1). The statistical analysis (Friedmans Two-Way Anova) indicated that significant differences ($p<0.05$) between the admission groups could be found in the total group (all study programmes) and at the Business administration study programme.

The results of the total group are shown in Figure 4, and indicate that the GPA-group shows the highest level of performance when compared to the other

groups. The SweSAT-group is followed by the SweSAT+WE-group which shows the poorest performance. The differences which appeared in the total group were located between the GPA- and the SweSAT+WE-group and occurred after all semesters except the first one.

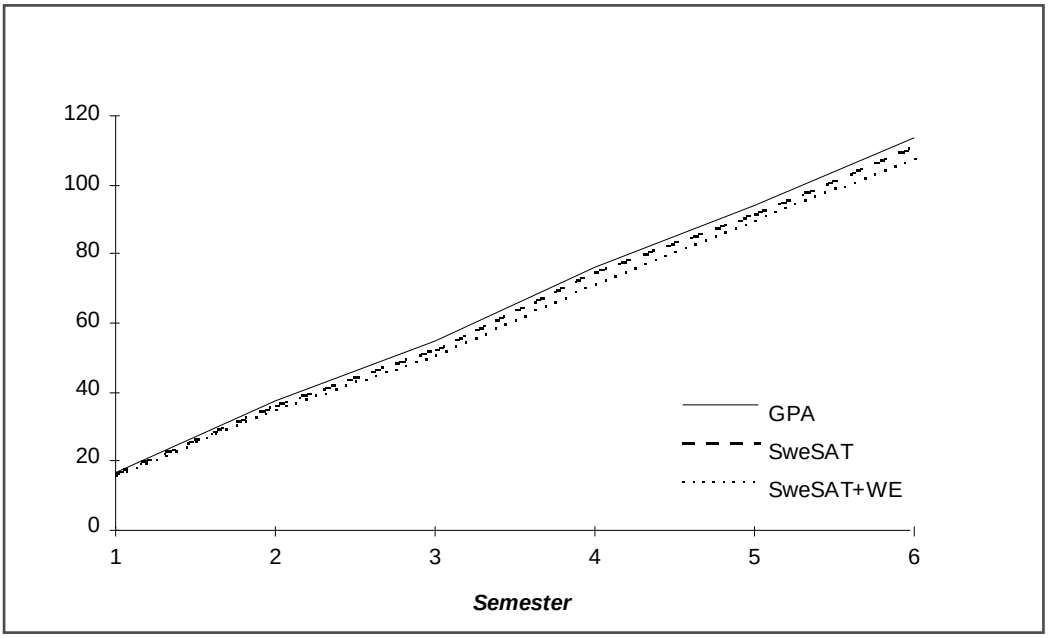


Figure 4. Number of credits acquired after each semester by the three admission groups. Remaining students. All study programmes.

The differences found in the total group could be related to the Business and administration study programme. Friedmans Two-Way Anova showed no significant differences within the study programmes, except for the admission groups at Business administration.

The drop-out frequency (Figure 3) after the second semester among students who had been admitted on the basis of SweSAT+WE at the medical study programme raises questions.

One question is whether students admitted on basis of their SweSAT+WE score were less successful in their studies at the Medical study programme. A closer look at Figure 5 reveals that this is not the case. According to this figure there are hardly any differences at all between students admitted on the basis of GPA, SweSAT and SweSAT+WE, respectively.

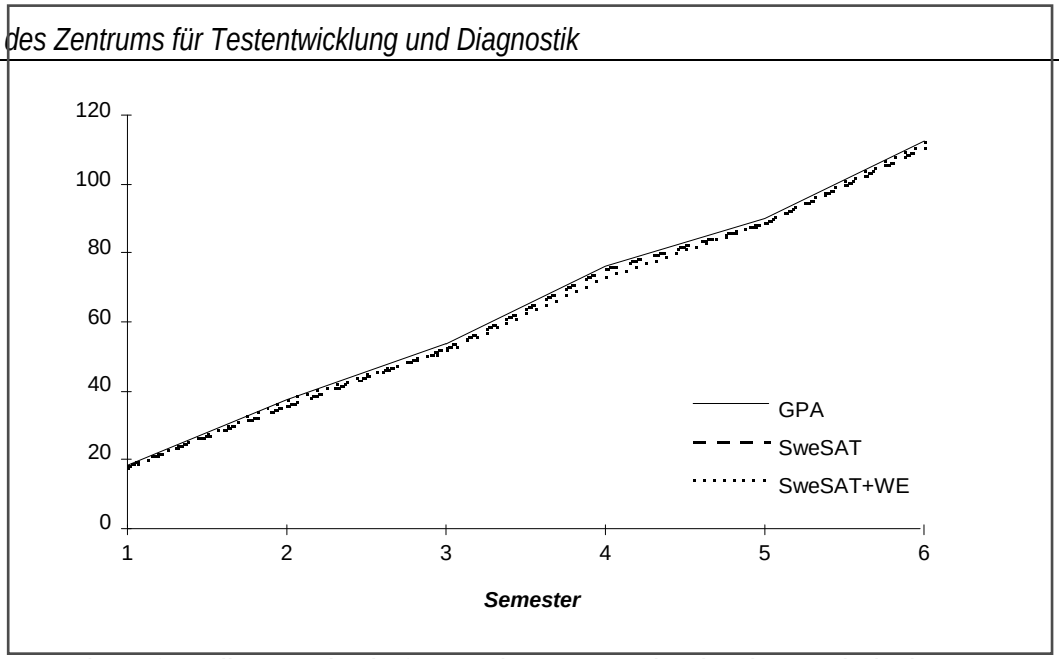


Figure 5. Number of credits acquired after each semester by the three admission groups. Remaining students. Medical study programme.

Another question to consider when analysing drop-out frequency concerns the reason for dropping out. In some cases, this has no relation at all to study success. For example, the cause ford dropping out may be that students change their field of study, that they stick to the study programme but move to another city, or that they get an employment.

The summarised and over-all conclusion from this study is that the selection procedure based on three admission groups, which has been used in Sweden since 1991, gives about the same results in terms of study success for students selected on the basis of GPA and on the basis of SweSAT-score. Maybe some questions could be raised about the success for students selected on the basis of SweSAT+WE at the Business administration study programme, but, on the other hand, the differences at the other study programmes were very small. At the medical study programme there were, for example, hardly no differences at all (Figure 5). It can also be added that the reason for giving merits for WE is the notion that students with WE in general give other students and the whole educational situation a broader frame of reference which is supposed to be fruitful in the academic setting.

3.5. Practical matters

Aside from the problems mentioned above, there are some miscellaneous issues that should be taken into account when introducing a test like the SweSAT Program. One of these is the issue of security. In our case this has lead to the development of a very strict procedure for printing, distribution, supervising the test administration, and for collecting and scoring the tests. The costs involved for these different measures have been quite large.

When a new version of the SweSAT is ready for printing, two staff members take it to the printing office which in our case is located a few miles from our office. The test version is handed over to the head of the printing office who then brings into a separate and locked room within the printing building. When the test is ready for distribution the official post service collects the tests and place them in sealed boxes for separate distribution to the different test centres all over Sweden (mostly universities). When the tests arrive there, a head test administrator (specifically trained for handling and administering the tests) takes over and locks up the tests.

At the test administration every applicant has to identify himself/herself and a seat is given to him/her by the head test administrator. At the end of the session, every answer sheet is checked against the applicant's ID. The answer sheets are sent to the computer department at our university where all sheets are checked against lists of the applicants participating at the different test centres. The process of scoring (optical scoring) then takes place in a locked room to which only the head of the computing process has access. In our case that means that as being chairmen of the Department of Educational Measurement and having overall responsibility for the test, we do not have access to the "scoring room" without permission from the head computing officer.

After the results have been obtained they are examined. Seemingly aberrant results are checked specifically and, if needed, then reported to the Agency of Higher Education that formally is in charge of the test.

4. Concluding remarks

Implementation of an admission program using a test battery like the SweSAT involves several components, some of which are more or less of a technical character. But there are also other components to pay attention to, among them political considerations. The story of SweSAT Program tells quite a bit of such political considerations.

Technically we know quite well how to develop an instrument to be used for selection purposes. In this respect we all can rely on a history of 100 years which started its second phase by the work of Alfred Binet and Henry Simon in 1905 (see for example Du Bois, 1970; Eswards, 1971; Linden & Linden, 1968). They proved, that by using a test consisting of 36 items it was possible to rank young pupils concerning their cognitive achievements. Since then remarkable technical progress has been made, but the basis for ranking students is more or less the same. The use of written tests of different kinds is still a very effective way of reflecting cognitive development and achievement.

The SweSAT Programme consists of five subtests of different kinds, ending up in 122 items in total. In principle it would be possible to use much fewer items

and still make the instrument effective for selection purposes. To do so would, however, make the test unacceptable for other reasons than technical ones. First of all it would be politically impossible, at least in Sweden, to do so. Such a test would not match the expectations of a serious instrument that covers a main part of what we call cognitive achievement is aimed at predicting future achievement. Secondly, the backwash effect would be tremendous, as students and their parents would request courses on such specific content as is reflected by the test. Other topics would soon be left out. On a more general level this is already the case, i.e. the content of an admission test will always has an effect on what is taught at study programs leading up to admission to further education. This, we think, is important to bar in mind when decisions about developing an admission test are to be made.

The arguments above quickly lead us on to the three critical elements in test development, namely:

1. The test has to be technically sound fulfilling basic requirements of reliability and validity

2. It has to match basic expectations of those taking the test, people at large and not the least politicians

3. It has to be cost-effective.

These three elements are not easy to combine. Test development is rather a question of balancing them. A test battery like the SweSAT is very much coloured by the efforts to do just that. We think that we have managed to develop a test battery that matches the three elements in a satisfactory manner. It's technical standard is very high (the easiest element to match), it has generally been very well received by the applicants and others interested in the admission process, including politicians (the test is evaluated on a continuous basis), and it is very cost-effective (the total cost per applicant is less than $25).

Programmes like the SweSAT are faced with a number of more specific issues than the ones just mentioned. The SweSAT Programme has (aside from a number of technical issues) been faced with primarily five issues, namely 1/social bias, 2/gender bias, 3/coaching, 4/prediction of academic performance, and 5/security matters. All these issues are examples of such considerations that have to be dealt with implementing a selection instrument like the SweSAT Programme. The importance of each one of these issues is probably dependent on each country's culture and other unique circumstances, but seen from a western world perspective each one of them has played a major part in the development of many testing programs in many countries.

The issue of social bias has to be treated seriously. The fact that a testing programme like the SweSAT will result in applicants from higher social groups achieving a higher score (on average) than applicants from lower social groups must not be seen as inevitable, i.e. every measure possible should be taken to limit the effect of social bias. Every item included in a selection test should be examined carefully to make sure that its content can be defended as fulfilling the basic aim of the test and not giving special advantages to certain groups of applicants.

The gender issue has been very much debated and researched. The measures mentioned with reference to social bias should be applied also in this aspect. It is a must that each individual item in a test be scrutinised as to the rational why it is "favouring" males (most typically the case) or females.

Coaching seems to be a never-ending story in the business of using tests for selection purposes. Even though the effects, according to the research that has been conducted so far, most often seem to be small we have to live with the fact that applicants might think otherwise. There are (at least) two sides of this coin. One is to make sure that the items and tests used are not coachable. Some item formats seem to be more coachable than others. In the process of developing the SweSAT Programme some 50 prototypes of item formats were treed out. Most of them were left out as they proved to be coachable. The other side of the coin says that is of uttermost importance that the applicants have a fair chance of getting acquainted with the item formats used. From the results of different coaching studies carried out so far, we know that applicants often improve their performance from the first time the test is taken to the second time and that this improvement most probably has to do with a better acquaintance with the item formats used.

The issue of prediction of academic performance has been with us since long ago in fact since long before Binet and Simon presented their test that meant a new start concerning how to measure cognitive achievement. Thousands of studies have been carried out in this field and tons of research articles have been presented. The results are generally the same in most of these studies. Grades and tests like the SweSAT Programme are the best predictors known for foretelling later achievement. Most of the variance in later achievement that is to be explained will be covered by these two instruments. Still, there is a must for every testing program to continue to conduct its own studies. The issue of prediction of academic performance exemplifies that the development of a testing programme for selection purposes involves a lot more than technical issues.

The security problem in any large scale testing program like the SweSAT Programme must be dealt with in a very cautious way. Every step in the process of

developing the test, printing it, distributing it, administering it, collecting the results and scoring it has to be carefully laid out.

Finally, a few words about the future of selection tests seen from a Swedish perspective. The SweSAT Programme was introduced at a time when there was a very critical atmosphere concerning every kind of individual given written tests. We sort passed that period with honour, i.e. the test seemed to be accepted by most persons, applicants, people at large and politicians. The 80's saw an even greater acceptance of the test. In the late 80's voices were raised concerning for broadening the set of instruments used for selection to higher education, and most colleges and universities were given larger freedom to develop their own instruments parallel to grades and the SweSAT Programme. Many of them took the chance of developing their own specific instruments, mostly interviews, portfolios of different kinds and more content directed tests. It is still too soon to evaluate these efforts, but up until today no results tell the story that a replacement of grades and the SweSAT Programme by these new instruments would guarantee a better prediction of academic performance. However, a two-step procedure might be a way to follow, where grades and results from the SweSAT Programme, constitute the first part with more specific instruments for the training program in focus as complements but again such a step has to be measured against the third element, namely cost-effectiveness.

References

Becker, B.J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. Review of Educational Research, 60, 373-417.

Du Bois, P.H. (1970). A history of psychological testing. Boston: Allyn and Bacon, Inc.

Edwards, A.J. (1971). Individual mental testing Part I: History and Theories. San Francisco: Intext Educational Publishers.

Henriksson, W. (1981). Effekter av övning och instruktion på testprestation. (The effects of practice and coaching on test score). Dissertation, Department of Education, University of Umeå.

Henriksson, W., & Wedman, I. (1992). Prediction of academic success in a per-spective of criterion-related and construct validity. Educational Measure-ment, No 2. University of Umeå, Department of Education, Division of Educa-tional Measurement.

Henriksson, W., & Wolming, S. (in press). Academic performance at four study programmes. A comparison of students admitted on the basis of GPA and SweSAT-scores with and without credits for work experience. To be published in Scandinavian Journal of Educational Research.

Henrysson, S., & Wedman, I. (1979). Research in Sweden with regards to predictive tests and interviews for admission to higher education. Spånor från SPINT, Nr 13. Department of Education, University of Umeå.

Henrysson, S., Kriström, M., & Lexelius, A. (1985). Meritvärdering och studie-prognos. Några undersökningar av antagningssystemets effekter (Evaluation of merits and academic performance. Studies of the effects of selection regulations). Arbetsrapporter från

pedagogiska institutionen, Umeå universitet, Nr 21, Pedagogiska institutionen, Umeå universitet.

Linden, K.W., & Linden, J.D. (1968). Modern mental measurement: A historical perspective. Boston: Houghton Mifflin Co.

Pike, L.W., & Evans, F.R. (1972). The effects of special instruction for three kinds of mathematics aptitude items. College Entrance Examination Board: Research and Development Report 71-72, No. 7 and ETS Research Bulletin 72-19. Princeton, New Jersey: Educational Testing Service.

Powers, D.E. (1986). Relations of test item characteristics to test preparation/test practice effects: A quantitative summary. Psychologi-cal Bulletin, 1986, 100, 67-77.

Stage, C. (1993). Gender differences on the SweSAT. A review of studies since 1975. Educational Measurement, No 7. University of Umeå, Department of Education, Division of Educational Measurement.

Vernon, P.E. (1960). Intelligence and attainment. London: University of London.

Wedman, I. (1992). Selection to higher education in Sweden. Educational Measurement, No 1. University of Umeå, Department of Education, Division of Educational Measurement.

Wedman, I. (1994). The Swedish Scholastic Aptitude Test: Development, Use, and Research. Educational Measurement: Issues and Practice 5-11.

Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L. (1990) Predicting college grades: An analysis of institutional trends over two decades. Educational Testing Service, Princeton.

The Medical College Admission Test (MCAT) - its use in U.S. and Canada and some results of validation studies

J. L. Hackett

Medical Education in the United States - Background Information

Medical education in the United States is provided by a mix of independent, private medical schools or colleges as well as medical schools supported or operated by the education system of one of the 50 state governments. There is only one Federally operated medical school, and it was established to train physicians for duty with the Armed Forces. There are 125 accredited schools of allopathic medicine in the United States. Successful graduates of these schools earned a Doctor of Medicine degree (M.D.). Other health professions institutions train students to become doctors of osteopathic, podiatric, or veterinary medicine. Most medical schools are university- based, and most require applicants to have successfully completed a course of studies in the sciences and liberal arts at an undergraduate, baccalaureate-degree granting college or university before they can matriculate to medical school. Many medical schools are associated with teaching hospitals at which medical school graduates receive their residency and medical specialty training following award of their M.D. degree.

In 1890, 66 medical school deans, motivated by a common desire to elevate the standards of medical education, met to form the Association of American Medical Colleges (AAMC). Initially, the primary focus of the Association was to aid the implementation of major reforms in academic medicine recommended by national studies, such as university-based medical education. The Association then turned its attention to improving the process of medical education. In the 1960s, the Association reorganized to support the full range of its members concerns -- education, research, and service to patients -- giving teaching hospital executives, medical school faculty, and medical students a voice in its governance. Today, the AAMC is a private, nonprofit association with its membership comprised of the 125 accredited U.S. medical schools; the 16 accredited Canadian medical schools; more than 400 major teaching hospitals; 86 academic and professional societies representing 87,000 faculty members; and the nation’s 67,000 medical students and 102,000 medical residents.

Services Provided by the AAMC

The Association of American Medical Colleges (AAMC) provides a number of services to its member medical schools and to medical school applicants and

students. The goal of these services is to promote quality in all areas of academic medicine. These services include:

The Liaison Committee on Medical Education (LCME). The LCME is the accrediting body for U.S. medical education programs leading to the M.D. degree. The American Medical Association (AMA), the professional association for practicing physicians, is a co-sponsor of the LCME.

MEDLOANS. MEDLOANS is a comprehensive loan program that provides financial assistance for enrolled medical students.

The National Residency Matching Program (NRMP). The National Residency Matching Program is a computerized program managed by the AAMC that matches candidates for training in medical residency programs at teaching hospitals throughout the United States according to the candidates' preferences and the needs of the residency programs. For 1996, there were 20,500 residency positions available in 3,400 specialty programs at participating hospitals. 24,700 medical school graduates participated in the Match. Of these, 8,100 were graduates of foreign (non-U.S. or non-Canadian) medical schools.

The American Medical College Application Service (AMCAS). AMCAS is a centralized medical school application service that enables applicants to file a single, standardized application with the AAMC. The AAMC reviews and verifies information on the application, including undergraduate college transcripts, course grades, and admission test scores, and forwards this standard data to any of the 110 participating U.S. medical schools to which the applicant wishes to apply. For the 1995 medical school entering class, 46,591 applicants applied for the 16,253 first-year positions in the 125 U.S. medical schools. Most of these applications were processed through AMCAS, and, on average, each applicant applied to 12 medical schools.

The Medical College Admission Test (MCAT). The MCAT is a standardized test used to assess medical school applicants' science knowledge, reasoning ability, and communication and writing skills. The MCAT was developed by the AAMC. The test is administered twice each year at 600 testing sites throughout the U.S. and Canada and at ten locations outside North America. During the two administrations in 1995, more than 67,000 persons sat for the examination. MCAT scores are one of a number of factors considered by medical school admissions committees at each school in evaluating and selecting applicants.

The Admissions Process at U.S. Medical Schools

Each medical school operates independently when evaluating applicant files and making admissions decisions. While most schools require applicants to apply

for admission by using the AAMC ‘s centralized application service (AMCAS), many schools also require that applicants complete and submit school-unique, supplemental application forms which provide additional personal data and information. The AAMC does not make admission recommendations or play a role in the admission decisions of the schools. Most medical schools have established admissions committees made up of faculty and staff members. These committees review applicants’ files and evaluate applicants’ qualifications based upon criteria developed by each school. Potential candidates for admission may be invited to come to the school for an interview.

This entire process may last a year. For example, a college student who plans to graduate from his or her undergraduate college studies in June of 1997 and who wishes to apply for a medical school’s entering class, which begins in September of 1997, most likely took the Medical College Admission Test (MCAT) in April 1996. Scores for this test were reported to the student in June 1996. If the student was dissatisfied with his or her scores, the student may have taken the MCAT again at its next administration in August 1996. The AMCAS and the medical schools which do not participate in AMCAS began accepting applications for the September 1997 entering class in June 1996. While the cut off date for receipt of applications vary by school, most schools have application receipt deadlines between October 15 and November 15. During the period from the receipt of applications until March 1997, admission committees at the schools evaluate applications and applicants, conduct interviews, and make final decisions. By March 1997, all applicants will have received an offer of enrollment from or been rejected by each school they have applied to. Some schools place qualified students, who have been rejected due to limits on enrollments, on a waiting list. Applicants on a waiting list may be offered admission in place of accepted applicants who choose to attend another school.

Admission testing has generated substantial public debate and many people have expressed concern about the fairness of standardized test, about test bias, and about over reliance on multiple-choice test scores in admission decision making. As will be indicated later in this article, data has shown that MCAT scores provide useful information about medical school applicants’ subsequent academic performance. A 1993 survey of medical school admissions officials revealed that MCAT data is used to:

- Identify applicants likely to succeed in medical school and those likely to experience academic difficulty,

- Assess applicants’ strengthens and weaknesses in knowledge of entry-level science content, science problem solving, critical and analytical thinking, and written communication skills,

- Interpret grade transcripts and letters of evaluation from unfamiliar undergraduate institutions.

Educational test and measurements experts tell those in the school admissions field that standardized test scores should never be the sole basis for student selection. The AAMC agrees with this philosophy and advises medical school admissions officials that MCAT scores are intended to be only one of several measures of an applicant's academic qualifications that they should consider. Many medical school admissions committees evaluate MCAT scores in conjunction with these other sources of information about the applicant:

- Undergraduate and postgraduate courses' grade point averages (GPA),
- Breadth and difficulty of undergraduate course work,
- Quality of the degree-granting undergraduate institution,
- Letters of evaluation and recommendation from undergraduate advisors, faculty members, and others,
- Involvement in extracurricular activities, such as student government and community service,
- Involvement in and quality of health related work and research experience,
- Participation in other activities demonstrating motivation, responsibility, maturity, integrity, resourcefulness, tolerance, perseverance, dedication to service, and other relevant characteristics,
- Medical school interview results, and
- State or county of legal residence in the United States.

The History and Development of the Medical College Admission Test (MCAT)

The AAMC first sponsored an objective test for applicants to medical school in 1930 for the purpose of reducing the high attrition rate among entering freshman. This test was called the Scholastic Aptitude Test for Medical School, and versions of it were used until 1946. At that time a new test, the Professional Aptitude Test, was introduced. In 1948, the name of this test was changed to the Medical School Admission Test (the MCAT). Since that time, the MCAT's content, format, and score reporting conventions have evolved and changed to meet the needs of medical schools to select and prepare students for the chang-

ing requirements of medical practice as the knowledge base and technologies of medicine rapidly change and expand.

The current version of the MCAT was introduced in 1991. It is the result of revision efforts started in 1983 with the formation of the Ad Hoc Advisory Committee to the MCAT Essay Pilot Project which was charged with determining the utility and feasibility of including a writing assessment as part of the MCAT. In addition, revision of the MCAT's multiple-choice test sections began in 1987, guided by the MCAT Evaluation Panel. Both committees were composed of AAMC constituents and staff. The two committees concluded their work by submitting a single set of recommendations which set forth their conceptual framework which would underlie the format of the 1991 revised test battery. These recommendations reinforced the medical community's view that physicians must be more than holders of vast amounts of scientific information. They emphasized the need for future physicians to have the ability to gather and assess data, to apply the basic concepts and principles of medicine to the solution of scientific and clinical problems, to evaluate situations critically, and to arrive at logical solutions. Physicians also must update their knowledge and skills continually and be able to communicate effectively with patients, colleagues, and the public. Mastery of basic concepts in biology, chemistry, and physics, while still considered prerequisite, was not judged to be a sufficient indicator of success in medical school.

Composition of the Current Version of the MCAT

The revised MCAT was designed to reinforce medical education's call for candidates with balanced undergraduate preparation and with knowledge of basic science concepts as well as the ability to think critically. Its three multiple-choice sections (Verbal Reasoning, Physical Sciences, and Biological Sciences) and directed writing assessment (the Writing Sample) were developed to assess (1) mastery of basic concepts in biology, chemistry, and physics, (2) facility with scientific problem solving and critical thinking, and (3) communication and writing skills. To discourage premedical students from concentrating too heavily in science course during their undergraduate education, science content coverage on the MCAT was restricted to the topics and subject matter taught in first-year biology, chemistry, and physics courses at most colleges.

The MCAT is composed of four distinct sections. The test is 5 3/4 hours in duration. Two sections are administered in the morning followed by the remaining two sections after a lunch break. Each section is scored separately and a distinct, scaled score is reported for each section. The sequence of each section and allocated times are as follows:

Section	Number of Questions	Time in Minutes
Verbal Reasoning	65	85
(break)		(10)
Physical Sciences	77	100
(lunch)		(60)
Writing Sample	2 topics	60
(break)		(10)
Biological Sciences	77	100

Verbal Reasoning

The Verbal Reasoning section of the MCAT is designed to assess the examinee's ability to understand, evaluate, and apply information and arguments presented in prose texts. The test consists of several passage, each 500 - 600 words long, taken from the humanities and social sciences and from areas of the natural sciences not tested on the MCAT Physical and Biological Sciences sections. Each passage is accompanied by six to ten multiple-choice questions based on the information presented in the passage. Some questions assess basic comprehension of the text; others require the examinee to analyze data, evaluate the validity of an argument, or apply knowledge gained from the passage to other contexts. Since the humanities, social sciences, and natural sciences include a vast range of subjects and since courses in these areas differ greatly in content, test questions do not cover a specific set of topics. Examinees are not tested for specific subject knowledge in the disciplines addressed by the Verbal Reasoning section.

The questions on the Verbal Reasoning test measure skills in the four cognitive areas of: comprehension, evaluation, application, and incorporation of new information.

Physical Sciences

The MCAT Physical Sciences section is composed of multiple-choice questions that assess an examinee's reasoning skills in general chemistry and physics. The majority of questions are based on passages, each about 250 words in length, that describe a situation or a problem. The test consists of ten or eleven of these problem sets, each followed by four to eight questions. An additional 15 questions are independent of any passages and of each other. Neither the

passage related questions nor the independent questions test the ability to memorize scientific facts. Rather, both types of questions assess the examinee's knowledge of basic physical science concepts and facility at problem solving in physics and physically-related chemistry areas. Many passages also include data presented in graphs, tables, and figures, and some of the questions related to these passages require interpretation of the data. The incorporation of items requiring examinees to understand and interpret data presented in graphs, charts, and tables with text material that crosses science disciplines provide a rich basis for assessing problem solving on the MCAT.

The concepts covered by the Physical Sciences section were determined by surveying undergraduate faculty members on the level of coverage given to potential topics in the introductory chemistry and physics course sequences. Medical educators, medical students, and physicians were asked the relevance and importance of potential topics for the study and practice of medicine. Chemistry and physics topics judged as prerequisite by medical school respondents and covered at a majority of undergraduate institutions in the introductory science course sequences were selected for inclusion in the test. The chemistry and physics concepts included in the test are considered basic. Advanced level undergraduate course work in physics and chemistry is not required to successfully take the test. A knowledge of calculus is not required. The mathematics concepts required for the Physical Sciences section are those typically required of students in introductory science courses. Topics included in the Physical Sciences test section include:

Acids/Bases	Atomic and Nuclear Structure	Bonding
Electrochemistry	Electrostatics and Electromagnetism	Electronic Circuits
Solution Chemistry	Equilibrium and Momentum	Light and Geometrical Optics
Fluids and Solids	Force and Motion, Gravitation	Phases and Phase Equilibra
Stoichiometry	Translational Motion	Sound
Electronic Structure and the Periodic Table	Work and Energy	Wave Characteristics and Periodic Motion
Rate Processes in Chemical Reactions	Kinetics and Equilibrium	Thermodynamics and Thermochemistry

Writing Sample

By requiring candidates to develop and present ideas in a cohesive manner, the MCAT Writing Sample offers medical school admission committees evidence of applicants' writing and analytical skills. These skills are critical to the preparation of useful medical records and to effective communications with patients and other health professionals. The MCAT Writing Sample provides unique information unavailable from other sections of the examination.

The Writing Sample consists of two essays, each of which the examinee is allotted 30 minutes to

write. The Writing Sample is designed to assess skills in the following areas:

- Developing a central idea
- Synthesizing concepts and ideas
- Presenting ideas cohesively and logically
- Writing clearly following accepted practices of grammar, syntax, and punctuation consistent with timed, first-draft compositions.

Each essay question provides a specific topic that requires an expository response. Essay topics do not pertain to the technical contents of biology, chemistry, physics, or mathematics; the medical school application process or the reasons for the choice of a medical career; social or cultural issues not in the general experience of MCAT examinees; or religious or other emotionally charged issues. Each of an examinee's essays is rated on a six-point scale by two separate readers. If scores for an essay vary by more than a point, a third reader reevaluates the essay. The total scores for both essays are summed and converted to an alphabetical scale for reporting. Medical schools receive the alphabetical score and, at their option, copies of their applicants' essays.

Biological Sciences

The format for the MCAT Biological Sciences section is similar to that of the Physical Sciences section. Multiple-choice questions are used to assess reasoning skills in biology and organic chemistry. The test includes ten or eleven problem sets with four to eight questions each and 15 questions independent of any passage and of each other. Passages are approximately 250 words in length. Some questions require interpretation of information found in graphs, tables, or figures. Questions do not assess rote memorization of scientific facts; instead they test knowledge of basic biological sciences concepts and problem-solving abilities in biology and the biologically-related areas of chemistry.

Like the concepts in the Physical Sciences section, biological sciences concepts were identified by surveying undergraduate and medical school faculty, medical

students, and physicians. Biology and organic chemistry concepts included in the Biological Sciences section are considered basic. They are taught at the introductory level at the vast majority of undergraduate institutions. Advanced undergraduate-level course work in biology or organic chemistry is not required for the test. The following topics are covered by this section:

Amines	Biological Molecules	Digestive and Excretory Systems
Evolution	Hydrocarbons	Generalized Eukaryotic Cell
Genetics	Skin Systems	Muscle and Skeletal Systems
Microbiology	Oxygen-Containing Compounds	Nervous and Endocrine Systems
Molecular Biology: Enzymes and Cellular Metabolism	Molecular Biology: DNA and Protein Synthesis	Organic Covalent Bonding
Respiration Systems	Separations and Purifications	Reproductive Systems and Development
Circulatory, Lymphatic, and Immune Systems	Use of Spectroscopy in Structural Identification	Specialized Eukaryotic Cells and Tissues

The MCAT provides admissions committees with a standardized measure of academic achievement for all examinees. Differences in undergraduate curricular emphases, evaluation standards, and grading scales make some preadmission information, such as undergraduate grades from different colleges, difficult to interpret. The MCAT provides assessment information about specific knowledge and skills on a standard scale for all examinees.

Test Construction and Discrimination Reviews

The test questions used in the multiple-choice sections of the MCAT (Verbal Reasoning, Physical Sciences, and Biological Sciences) are designed and written individually by college and university professors from throughout the United States. These are experiences teachers in the fields of chemistry, physics, biology, and the humanities who have contracted with AAMC to confidentially develop and write MCAT test items. Each item writer is assigned a specific subject area and skill level which his or her test passages and questions must examine. Submitted questions are initially reviewed by our in-house editors. The test questions are then forwarded to external subject matter experts for

technical review and to prominent and educators who examine the questions for potential gender and racial or ethnic bias or discrimination.

After test questions have been thoroughly reviewed, edited, and approved for inclusion in the MCAT, each passage and its related questions are pre-tested as field test questions. In each operational MCAT test form which is administered, at least one passage and its related questions is a field test question set. These questions are not scored. Test takers know that such questions are included in their test, but they do not know the position or location of the field test questions. Subsequent to a test administration, each test question undergoes intense statistical analyses. Degree of difficulty for each question is calculated based on student performance on the question. Differential results by students' gender and racial/ethnic backgrounds are examined to identify any previously undetected question bias

Following all these reviews and analyses, passages and questions are approved for inclusion in an operational MCAT examination as scorable test items. Questions are assembled into operational test forms based on the subject matter and skill levels each question assesses and on each questions degree of difficulty as determined during field testing. This process ensures that each test form assesses the knowledges and skills required by the MCAT Test Specifications and that different versions of the MCAT have comparable levels of difficulty.

Predictive Validity Research Study for the MCAT

Concurrent with the introduction of the revised MCAT in 1991, a plan for assessing the predictive value of the new MCAT was designed. Fourteen medical schools were selected to participated in the study, and they agreed to provided the necessary data. These schools were chosen to represent the 125 U.S. medical schools. The 14 schools are located in different regions of the country; some are privately operated and the remainder are state supported; the selected schools have a proportional mixture of different types of curricula; and the size and racial or ethnic composition of their student bodies are representative of most of the other U.S. medical schools.

The study is following two cohorts of students comprised of those who entered the selected medical schools in the fall of 1992 and in the fall of 1993. In the U.S., the academic school year at most schools begins in August or September and ends in May or June. Students in these two cohorts were identified, and preadmission data (data which was available to the medical schools' admissions committees when their admission decisions were made) were gathered on the students. Preadmission data obtained for each student consisted of undergraduate grade-point average (UGPA), MCAT scores, and the selectivity index for

their undergraduate institution. The undergraduate grade point average is a numerical grade used in the U.S. which is the average of all of a student's college course grades weighted by course credit hours. The selectivity index is a quality index number assigned to all undergraduate colleges and universities in the United States based upon the mean college admission test scores of their students enrolled in a given year. These data allowed evaluation of the MCAT's predictive value within the context of other data typically used during admissions decision making.

To date, performance data for the 1992 entering class at the selected medical schools has been gathered. The students' performances were summarized by computing end-of-year grade-point averages (GPAs) for each year (year 1 of medical school and year 2 of medical school) and a two-year cumulative grade-point average which was the simple average of the two end-of -year averages.

The extent to which MCAT performance predicts performance on Step 1 of the United States Medical Licensing Examination (USMLE) was also examined. The licensing exam is administered in "steps" during the course of a medical student's education and training. All three Steps of the USMLE must be passed before a medical school graduate is licensed to practice medicine in the United States. Step 1 of the USMLE is taken by most medical school students during their second year of medical school.

Before analyzing the relationships between preadmission variables and medical school grades, descriptive information for each of the preadmission variables were obtained. The data were not collapsed across the schools; instead, the distribution of each of the preadmission variables was investigated within each of the schools.

The relationships among the MCAT, Undergraduate GPA, institutional selectivity, medical school grades, and Step 1 scores were investigated using multiple regression analysis. Regressions were run separately for the entrants to each school. Because applicants with low values on preadmission variables are typically not selected for medical schools, the range of values for preadmission variables is generally restricted or truncated. Corrected correlations estimate the strength of the relationships between the preadmission data and the medical school performance variables in the absence of selection. The restriction-in-range corrections were based on MCAT, UGPA, and selectivity data for each school's applicant pool.

Multiple correlation coefficients were obtained for five unique predictor sets, and the results were summarized across schools. The five predictor sets were:

Undergraduate GPAs

MCAT scores

Undergraduate GPAs + selectivity index

Undergraduate GPAs + MCAT scores

Undergraduate GPAs + selectivity + MCAT scores

The following table reports the ranges and median values for the corrected multiple correlation coefficients obtained for the five sets of predictors. As is evident in the table, MCAT scores appear to have a slightly higher correlation with medical school grades (median correlations ranging from .62 to .67) than do UGPA data (median correlations ranging from .54 to .58). However, prediction of performance was improved when the two sets of predictors were considered jointly (median correlations of .70 to .76).

Range of Median Values of the Corrected Correlations between Preadmission Data Sets and Medical School Performance for the 1992 Entering Class at 14 Schools

	Medical School Performance			
Predictor Sets	Year 1 GPA	Year 2 GPA	Cumulative Year 1&2 GPA	USMLE Step 1 Score
Undergraduate GPAs				
Median	.54	.58	.58	.48
Range	(.40 - .74)	(.23 - .74)	(.42 - .73)	(.35 -.64)
MCATs				
Median	.67	.62	.64	.72
Range	(.38 - .78)	(.46 - .72)	(.44 -.78)	(.58 - .79)
Undergraduate GPAs + Selectivity				
Median	.60	.61	.64	.53
Range	(.50 - .80)	(.29 - .79)	(.49 - .82)	(.40 - .73)
Undergraduate GPAs + MCATs				
Median	.75	.70	.76	.75
Range	(.48 - .80)	(.61 - .77)	(.63 - .81)	(.62 - .81)
Undergraduate GPAs + MCATs + Selectivity				
Median	.75	.73	.76	.75
Range	(.53 - .82)	(.64 - .81)	(.63 - .84)	(.62 - .82)

A substantial difference is evident when comparing median correlations for MCAT and UGPA data with USMLE Step 1 performance. Here, MCAT is much more strongly related to Step 1 (median value of .72) than UGPA (median value of .48). Again, however, when MCAT and UGPA data are considered jointly, the median correlation coefficient increases.

The MCAT's utility in the admission process becomes evident when values obtained for predictor set 3 (UGPA + selectivity) are compared with those ob-

tained for predictor set 5 (UGPA, selectivity, and MCAT). The median values of the prediction correlations increase notably when the MCAT scores are added to the equation.

Because the MCAT's primary purpose is to identify individuals most likely to succeed in medical school, initial efforts have focused on the evaluation of the predictive validity of the test. This study to date represents a comprehensive evaluation of the predictive validity for the first two years of medical school, which in the United States are heavily weighted towards the study of the basic sciences. Establishing the predictive validity of the MCAT for these first two years of basic medical science study is a crucial step in the overall validity studies planned for the MCAT. Data are being collected from the entering class of 1993 to replicate the present study's results for the 1992 entering class. Subsequent studies will examine correlations with performance in the clinical years of medical school training as well as performance on the other, sequential Step Examinations that comprise the U.S. Medical Licensing Examination process.

Summary

The decision to select and to admit an applicant to a medical school in the United States is one that is made individually by each medical school according to each school's own criteria. Multiple factors are reviewed and considered by medical school admission committees when making admission decisions. One of these factors is the applicant's performance on the Medical College Admission Test (MCAT). The current version of the MCAT was specifically designed to access the skills identified by medical school faculty and physicians as those necessary for a student to succeed in medical school and in the practice of medicine. The results of MCAT predictive validity studies conducted to date indicate that the MCAT is a strong predictor of performance in the critical first two years of medical school. Additional studies are planned which will extend the examination of the MCAT's correlation with performance into the third and fourth years of medical school students' clinical studies and to their performance on the final stages of the licensing exam process.

References

Material in this paper was adapted or extracted from the following sources:

Leadership for Academic Medicine. Washington, DC: Association of American Medical Colleges, 1996.

MCAT Student Manual. Washington, DC: Association of American Medical Colleges, 1995.

Use of MCAT Data in Admissions: A Guide for Medical School Admission Officers and Faculty. Washington, DC: Association of American Medical Colleges, 1994.

Mitchell, K., Haynes, R., and Koenig, J. Assessing the Validity of the Updated Medical College Admission Test. Academic Medicine 69(1994): 394-401.

Wiley, A., and Koenig, J. The Validity of the Medical College Admission Test for Predicting Performance in the First Two Years of Medical School. Academic Medicine 71 (Oct Suppl. 1996): S83-S85.

Admission to the study in medicine in Belgium: two 'different' solutions to the 'same' problem; reflections of a Flemish school psychologist

P. J. Janssen

A few weeks ago a specialized Flemish magazine for general practitioners (the so called 'Artsenkrant') announced that Belgium at this very moment disposes of 35.000 family doctors and medical specialists. That means an average of one medical doctor for every 280 inhabitants. Annually about 1200 students complete their seven years of medical basic training; from the year 2004 only 700 of them will be allowed - as was decided by the federal ministry of National Health about a year ago - to enter the profession. The problems in the domain of dentistry are analogous and need to be solved in the same time. That's the problem my country has to solve to the best of its possibilities, being obliged to do so for several reasons: to control the costs of the national health insurance, to maintain the quality of medicine and dentistry on an accepted level (as the organizations of practitioners were claiming). Due to the way the Belgian State internally is organized - now going its way into a federation - this 'central' decision has to be elaborated within each of it's both 'national' governments, responsible for their educational systems as such. That explains way this 'same' problem - within 'one' but subjectively not 'the same' country - could create two 'different' solutions.

In order to enable full understanding of this problem from the Belgian perspective, I start my contribution by presenting some additional information about my country (1). Next I will present the two solutions for the problem as they will be elaborated from now on (2). Being personally involved in the development of the selection procedure, as it will be applied in the Dutch speaking part of my country, I will present you the arguments of the school psychologist I am, for preferring this Flemish approach (3).

1. The Belgian context

As many of you know, Belgium is from its origin in 1831 a parliamentary democracy with a constitutional hereditary monarchy. The governmental system set up by the constitution of 1831, which vested most powers in the central government, was modified by the division of the country into linguistic areas in the 1960s. Constitutional revisions in the 1970's and 1980's resulted in the establishment of a quasi-federal governmental structure based on the three official linguistic communities: Dutch (about 57 % of population), French (42 %) and German (1 %). These got their own 'national' governments and councils, which acquired a considerable degree of influence over regional and national affairs.

One of their domains of autonomy refers to education. One of the implications was already mentioned: the 'national' problem of too many medical doctors will be solved by at least two different approaches.

To understand the way the problem is posed, one also has to know that nearly all Belgians are covered by national health insurance, which as such is, at least until now, managed for Belgium as a whole, thus on the federal level. This social security system also covers work accidents, unemployment, premature death, occupational diseases, invalidism, and retirement pensions. The financial cost of such a system is enormous. The number of medical doctors is one of the system's parameters, which have to remain controllable. Now one understands why the central level could formulate a 'problem' that needs to be solved on both national sub-levels...

One of many other common Belgian problems refer to the transition youngsters make from secondary to tertiary education. The latter implies university as well as the so called Higher Education outside university. Both subsystems are freely accessible for every youngster who has completed her and his general secondary education. There are only two exceptions: the one refers to the five year university training in engineering (within faculties of applied sciences); the other to some fields within the artistic Higher Education outside universities. For both a freshman has to pass an entrance exam assessing competency, resp. in mathematics or artistic potential. Everyone who passes, has the right to register as a student. Until now there is no numerus clausus or fixus. Important to know also is that Belgian universities are subsidized by their government on the base of enrollment figures: the more students (and the more freshmen), the 'better', at least at the beginning of the academic year... Also important to know is (1) that Belgium as such has no common 'national' achievement test or exam (some type of 'Abitur') at the end of secondary education; institutions themselves (also subsidized on the base of enrollment numbers !) autonomously award the degree and ...(2) that there exist real differences between these institutions, also as far as educational levels attained. The consequence of this quite 'complex' and at the same time quite 'generous' system are, at least, twofold:

(1)The first refers to the quite low success rates at the end of the first year, where the student - as he has to do every year during the curriculum - has to pass a fixed number of exams, each referring to a course he is obliged to take. Decisions are taken by a jury, grouping the professors who taught these courses. The student who is unsuccessful at the end of June, may try again at the end of August. The success rates at the end of the first year - both exam periods combined - vary from 1 out of 3, to 1 out of 2. Relatively most successful are students in engineering (about 65 %), immediately followed by students in medicine, what proves, as a matter of fact, that on the average this faculty already attracts capable and well motivated students.

(2)The second consequence has to do with the attractivity of this system of free entrance for young Europeans, who, being not allowed to the study of their choice by restricted entrance regulations within their own country, detect Belgium (as 'the land of hope and glory' ?) where they can start the study they want. So Flanders already for many years 'had to accept' and 'to pay' for many hundreds of students from the Netherlands, who were eliminated from medicine by the system of lottery in their own country. Some years the Dutch freshmen even outnumbered their Flemish colleagues in the Antwerp university ...

The basic curriculum in medicine counts 3 years of 'candidature' plus 4 years of second cycle. The study program of the very first year enables freshmen to master the necessary basis in mathematics, physics, chemistry and biology. During the second and third year anatomy and physiology are extensively studied. During the second cycle students get their specific medical basic training. Belgian medical faculties have their own educational climate. At the end of these 7 years of study students are 'selected' for specialization on the base of study results obtained during the preceding six years. This rule may explain why students in medicine prove to be hard working, and most of the time quite competitive, which in turn may explain some 'negative' behaviours of some students with respect to materials - for instance in libraries - which also might be of interest to their fellow-students. This 'third cycle' of specialization varies in time; it can take up to 5 years.

In 2004 the federal government will allow only 700 new medical doctors to enter the national health system. That year was chosen in order to offer both national communities sufficient time to take the measures they judge essential. In the case this measure had to be applied immediately, about 500 'doctors' out of the 1200 as 'delivered' this year, had to find another job ...

2. Different solutions to the 'same' problem

That is the problem my country was confronted with about a year ago. The two national Ministers of Education - Grafé in the French-speaking and Van den Bossche in the Flemish community - had to develop an option. In the case one would apply a division according to the percentages within the population, this would imply a reduction to about 399 doctors in Flanders and 301 in Wallonia . Different measures can be taken: one can introduce them at the end of the curriculum, apply them at the moment of transition form the first cycle into the second (after three years of study), take them at the end of the first year, or do it before students can register. Within each community discussions started immediately.

2.1. The 'French' way of solving (or postponing ?) the problem ...

In the French-speaking part the decision was taken to postpone this selection to the end of the first study cycle. So at least everyone, knowing what will happen, may try his chance. After three years only a limited number of students will be allowed to continue in medicine. The others - if there are more students than needed - have to find another study domain. For the moment it is not clear - at least not for me - how these future medics will be selected. It can happen on the base of their study results; in that case, I guess, one needs some type of national test, because these students study in different universities (Liège, Bruxelles, Louvain-la-Neuve). Another possibility would include that they, as 'candidates in medicine', have to apply for some type of a lottery system.

2.2. The solution in Flanders

In the Dutch-speaking part of Belgium the discussion about measures to be taken quite soon became dominated by the decision taken by the Minister of Education, Luc Van den Bossche, member of the Flemish Socialist Party. He decided to introduce a 'national' entrance selection before the beginning of the academic year 1996/1997. Every youngster wanting to study medicine has to take part; only those who will succeed, will be allowed to choose a university and to register as a freshman. A working group was charged on December 1, 1995, with the task to design and implement a procedure, which should be constructed in such a way, that future students will not be selected on the base of their educational attainment in Secondary Education, but 'only' on the base of their real medical aptitude, including also their motivation. ...

2.2.1. The procedure as developed in the beginning of 1996

This working committee consisted of 21 members, experts (from the Ministry of Education, university professors lecturing in medicine and psychometrics) as well as representatives of students in medicine. These people met about 10 times under the dynamic direction of Jan Adé, the Director of Higher Education and Scientific Research in the Flemish Ministry of Education. On January 26 the commission could meet Dr. Günter Trost, who described the German TMS and the way selection for medicine is functioning in Germany.

In the beginning of May 1996 - at the end of this same month the Flemish Parliament had to decide about a pro-position concerning the introduction of this entrance test - the Commission presented its report to Minister Van den Bossche. This document describes the work done and motivates the options chosen after a thorough evaluation of the pro's and con's of all kinds of measures. Interviewing of all candidates - about 1000 youngsters are expected to take part in the entrance examination - turned out to be for more than one reason completely impossible. The same holds for the so called 'apprenticeship in

nursery', enabling a short period of work in a hospital, which the representatives of the students judge to be an essential component of the procedure in order to enable a - at least subjectively - valid test of a youngster's motivation for the medical profession. Even a test for manual dexterity had to be excluded for students in dentistry, who also had to participate in this entrance examination.

The resulting test will consist of two big parts, each taking a full 'day' of testing. (1) In the first 'Knowledge of and Insight in Sciences' (KIS) will be tested by means of four 'different' instruments, resp. assessing an entrants competency in the domains of mathematics, chemistry, physics and biology. In each of these the content will be defined in terms of the minimum program Secondary Education is offering; the items will be constructed in such a way that they 'test' productivity - i.e. the constructive use of knowledge and insight - rather than pure reproduction. These tests proved to be essential, because Belgium does not, as other European countries do, test achievement at the end of Secondary Education by means of an uniform and common final exam. (2) During the second day an applicant's capability in the Acquisition and Elaboration of Information (the AEI-part of the exam) will be tested. Here instruments have to be developed to measure inter-individual differences con-cerning cognitive functioning, memory, assimilation of visual information and last but not least the ways these candidates handle a casus, describing, by means of modern audio-visual media, the complexity of a real-life medical problem from the perspective of the patient as well as the 'different' medical experts involved on the base of the information each of these disposes of. The idea for this last test was developed by my colleague Prof. Paul Coetsier from the university of Gent on the base of an expertise already available in the domain of selection of managers for industry.

Given the fact that the Flemish Minister of Education initially obliged the commission to organize this entrance exam - as a first try-out - twice before the start of the academic year 1996/1997, the commission simultaneously had to develop a strategy concerning the decisions to be taken on the base of the results of this two-days testing program. As a consequence two decision rules were formulated:

a) The first reflected the criteria for grading as they are used in Flemish universities. Everyone who reaches the criteria put forward, will receive at least a score of 10 out of 20 points. This score will be as higher as the level on which this performance can be placed, exceeds these minimal objectives; as such a student can receive distinction, great distinction or the greatest distinction, in the case his level of attainment reaches a core of 14/20, 16/20 or 18/20. Given the fact that not all parts of KIS are equally well taught in Flemish Secondary Education, the commission decided to take only each applicant's three 'best' scores into consideration. As far as AEI is concerned the same criterion holds.

By consequence, in order to succeed in this entrance exam, an applicant has to reach in each of the two parts of this entrance test at least a score of 10/20.

b) The commission decided also to allow - given this very first application of the procedure - also those youngsters, who, after the combination of their main scores on the two parts of this exam, obtained at least a T-score of 40 win the total group of applicants. This decision would imply that at this very first try-out of the entrance procedure 'only' about 15 % of the applicants would be not successful at all. So quite a safe start could be made, given the political considerations that urged Minister Van den Bossche to take this measure.

In the beginning of June the Flemish Parliament accepted this proposal. At the same moment, the execution of the decision taken was postponed until the start of the academic year 1997/1998 ... The reason was quite acceptable; it turned out to be too late to inform last year students in Secondary Education about the implications this decision would have on their vocational planning ...

2.2.2. Procedure as to be applied in 1997

In the same time it became clear that the first application of the parliamentary decision will take place at the be-ginning of the next academic year. In the way this Flemish parliamentary decision was taken, it implies some and even important changes from what the working group proposed to the Minister with respect of a very first application before the start of the academic year 1996/1997. The idea of a two days exam will hold. There will be, according to the Belgian academic tradition, two exam periods, implying so a second chance for those who did not succeed at the first occasion. Not well defined are both points in time involved: From a traditional university point of view the beginning of July and September seem appropriate; taking the perspective of a school psychologist, by definition well informed about the ways last year students in Secondary Education are going through their process of vocational decision making, two other moments seem to be more appropriate: the period after Eastern and the end of the school year in the beginning of July. The minister has to take his decision in one of the forthcoming weeks.

Between times it became clear that in the decision, as it was taken by the Flemish Parliament, the norm to succeed has been changed drastically as compared to the one proposed by the initially installed working group. In order to succeed, an applicant has to obtain now for each of the two constituent parts a score of at least 12/20. In the forthcoming days the jury - under the direction of chairman Jan Adé of the commission that prepared the procedure - officially will be installed. From that moment on, the way the entrance exam, as it will take place within the Flemish Community, will get it's definite content and form. It will be organized by an already existing governmental service - the

'Permanent Recruitment Secretariat' ('Vast Wervingssecretariaat') - which disposes of the necessary material and logistic infrastructure.

3. A school psychologist's evaluation

For the moment I see one essential aspect of the problem, which as such has not yet appropriately been analyzed. Until now the Flemish Minister Van den Bossche refuses to define the contingent of students to be allowed to start next year their studies in medicine. This means an extra complication for the commission who will be in charge of the preparation of the entrance examination and for the jury which has to decide about each individual applicant.

The idea of an entrance examination as such is in my view a sound approach towards the solution of a problem which - at least as I see it - is broader than this specific case in medicine and dentistry. As already mentioned, Belgium is confronted with quite low, even too low, success rates at the end of the first year in Higher Education. At least as far as Flanders is involved, I myself already proposed in 1988 (Janssen & De Neve, p. 179 - 183) a set of measures, which some years later was taken over by a Flemish group of experts in university education (the so called 'Contactgroep Academic Onderwijs') (CgAO 1993). These include the improvement of the ways last year students in Secondary Education have to be counseled in the making of their vocational decision and the didactic of the first year in university. Included within this counselling process are two series of tests, to be taken individually at two different moments during the process of vocational decision making. The second one could be replaced by an entrance selection, as it is proposed here for the studies in medicine and dentistry.

The proposal itself is based on the two big causes for these transition problems, as both are identified now on the base of about half a century of scientific research. The one refers to the lack of competence of entrants in Higher Education (3.1); the other to the quite defective self regulation of freshmen in Higher Education (3.2). Although both problems turn out to be, to a certain degree, independent (even 'competent' students do not succeed), they can be solved together by means of, as was already said, the improvement of student guidance during the last year in Secondary Education (3.3).

3.1. Competence: necessary but as such insufficient

An analysis of the process of studying makes clear that a serious lack of competence creates a break-down in the progression of studying as a process. As such studying has to be defined (Janssen, 1996) as the integration of learning and thinking. Both components can be described separately; so it becomes quite clear what their integration has to imply.

Learning has to be meaningful and comprises, according to Ausubel (1963), within a positive feedback loop the three stages of assimilation, transformation and accommodation. Assimilation implies complete understanding - on the base of available foreknowledge - of what is presented in lectures. Transformation implies intelligence, because the student himself has to change his until then appropriate cognitive structure into a new logically coherent 'Gestalt', which will allow him, having put this new one into memory, accommodation, i.e. the expertise to solve the problems he will be confronted with as the expert he wanted to become.

The impact of **thinking** was already mentioned in my description of the transformation the student, as a learner, has to achieve. But it also works, as can be seen within a moment, at both other stages involved. Thin-king can be threefold, resp. in line with the so called combination, comparison and cause-effect schemes. The first implies the bringing together of two separate elements into one new whole, the second the understanding of something new on the base of the 'old' one already is familiar with, the third the 'working' of one element in the 'making' of another ... These three - as was demonstrated by Gordon Pask (1976) - cannot work independently from each other, given the complexities students are confronted with.

Integration of the three stages of learning with the three ways of thinking creates - as is shown in figure 1 - the going together of up to nine 'different' mental operations, all involved within the process of studying. The loop in learning as well as the interdependencies in thinking remain 'working' within this 3x3 matrix scheme. So up to four loops can be identified within a psychologically well cohering way; each of which - at least in principle - can be tested, also before entering into a specific domain of study in Higher Education: (1) So cognition and memory have to enable the student by means of selective comparing the understanding of the novelties - in mathematics, physics, chemistry, ... - he will meet within a specific study program. (2) By integrating convergent production (being the systematic and quite logic approach of a problem) and 'selective combining' one can test a students capacity in problem solving, reasoning (as such the factoranalytically identified dimension 'reasoning ' within the German TMS)... (3) By integrating 'divergent production' (a person's capacity to produce meaningful 'wholes' in making a synthesis) and 'selective encoding' (or that same person's capacity to analyze a complexity by distinguishing the relevant form the irrelevant elements within a given context) one can test a student's capacity for finding meaning within so far new information (it remembers me the dimension 'visual information processing' within the German TMS) ... (4) By integrating organizing and timing, as such quite important for a student's self regulation, one can test in advance a student's capability in a fourth essential component of his studying as a process (as such the psychological meaning of the 'factor' memory as identified within the German

TMS ?) . (+1=5) In the meantime the operation ‘evaluation’ turns out to be the real hinge in this same studying; it refers, as will described immediately, to the way a student is capable to integrate these different process dimensions into (his experience of) ‘optimal functioning’ in what he is undertaking.

STUDYING as experienced as a students behavior	to try: What ? Why ? Meaning	to can: to be able to	to try: How hard ? Effort
Person	Intention	Self confidence	Activity to control
Task over time	Agenda	Effectiveness	Discipline
Environment	Relevance	Comfort vs difficulty	Impact, demand

Figure 1. The nine intellectual operations as involved in the process of studying, implying the integration of the 3 stages as involved in learning (from assimilation into accommodation and backwards) and the 3 ‘schemes’ of thinking, implying the going together - by means of a student’s permanent evaluation - of 4x2 mutually well co-ordinated mental activities.

There is an *alternative way* to test a student’s competence for what he is planning. Instead of confronting him with a series of specific tests, one can offer him the opportunity to study, during at least two hours, a representative sample of study materials form the domain he is interested in. This has the big advantage of offering him a real - at least in his eyes - face-valid test or try-out. He probably will be more inclined to accept the conclusions resulting from that ‘test’ or ‘personal experience’ as to be deduced from it. In line with that principle Minnaert & Janssen (1992) constructed a predictive as well as a nomological valid ‘Excursion into psychology as the domain of my future study’, explaining up to 56 % of variance within study results after a five year follow-up period. Such a type of instrument resembles quite well the casus my colleague Coetsier is constructing for the Flemish entrance test. In the meantime it became quite clear that the student’s study behaviour itself is an as essential component as competency itself turned out to be ...

3.2. Self-regulation: necessary but insufficient, when ...

This study behaviour has to be defined - so the internal logic of figure 2 can immediately be understood thanks to one’s "capacity for finding meaning" (as an aspect of ‘visual information processing’) - as the task-oriented interaction between the student (as a person) and his environment (or the study landscape

he is travelling in), enabling him the achievement of the goal set forward and naïvely attributing its attainment (according to Heider (1958) as the founding father of modern psychological theories of causal attribution) to the product of ‘to can’ (or ‘to be able to’) and ‘to try’ (combining so the ‘what and why ?’ or meaning with the ‘how hard ?’ or working in his doings). So up to nine specific study experiences - as identified in the mean-time by means of factor analysis of item materials of Likert-type study behaviour questionnaires - can find their place within an analogous 3x3 matrix concept. Each of these nine can be psychologically measured in a quite reliable way, offering so the individual student a real life picture of his study method, i.e. his - whether or not - effective regulation of his ‘way-into’ the criteria he has to meet at the end of the year ...

Divergent production (GUILFORD)	**Cognition / Memory (GUILFORD)**	**Convergent production (GUILFORD)**	**Accommodation (Doing; expertise)**
Organizing	**Evaluation (GUILFORD)**	**Timing**	**Transformation (doing into knowing and vice versa)**
Selective combining (STERNBERG)	**Selective comparing (STERNBERG)**	**Selective encoding (STERNBERG)**	**Assimilation of novelty (knowing)**
thinking by means of the combination scheme	**thinking by means of the comparison scheme**	**thinking by means of the cause effect scheme**	**STUDYING as a process:** product **of Thinking x Learning**

Figure 2. The nine experiences as involved in studying, constituting 4x2 'loops' (each reflecting a specific type of motivation), their going together resulting in effectiveness, intrinsic motivation or 'deep level learning', as such enabling the student the self-regulation of as many operations as involved in his studying as a process.

This effectiveness has to be psychologically understood as intrinsic motivation, implying that the student - whether or not - is experiencing his studying as rewarding in itself. In the positive case he *reaches 'deep level learning'* (Marton & Säljö, 1976), as such the essential condition for the development of real expertise. In that case this very same intrinsic motivation integrates up to four 'different' but all essential motivational subsystems for his optimally functioning as a human being: **(1)** *competence* motivation (the effectively going together of 'self confidence and 'situational comfort'); **(2)** *causality* motivation (the effectively going together of 'activity to control' and environmental 'relevance'); **(3)** curiosity or *creativity* motivation (the effectively going together of 'intention' or personal interest and situational 'demand') and **(4)**, last but not least, *stability* or self regulation motivation (the effectively going together of 'agenda' or time perspective and 'discipline').

In the meantime the parallelism between figures 1 (operations involved) and 2 (the complementing study experiences) merits further attention. So, indeed, the process of studying (figure 1) can be regulated by the individual student on the base of his *experiences* (figure 2); so the 'central' experience of effectiveness, as such resulting from the permanent evaluation of the ongoing process, becomes the necessary and sufficient situational knowledge, which enables the student, so becoming a studax or real expert in studying (Janssen, 1996), to be(come) and to remain effective in Higher Education. *'Only' three conditions have to be fulfilled:* **(a)** a sufficient expertise in studying (which can be achieved by

everyone who before has 'learned to learn' during secondary education), **(b)** a sound vocational decision (as such resulting into an internally well coherent matrix of the study experiences to come) and, **(c)** last but not least, a sufficient ... competence (as described in the preceding paragraph).

Both last conditions imply each other to a certain degree. Indeed, it is quite probable that the person who is making a sound vocational decision, will detect a potential conflict between his competence and the intellectual requirements as involved in a specific alternative, in due time and will take the consequences quite seriously... He spontaneously will change his planning towards another alternative offering him the highest 'value for money', i.e. the highest guarantee for real effectiveness. So this 'hecatomb' in the transition from Secondary towards Higher Education - at least in Flanders - can definitely be cleared up: *One has to 'organize' the counselling that last year students in Secondary Education need, in such a way, that each of them is capable to take into consideration his strong as well as his weak competencies before, or at least while, making his vocational decision.* A careful analysis of that vocational decision - as such it turns out to be a process, during which up to five stages have to be passed through - reveals the necessity *to offer these youngsters the right information at the right moment, so enabling them to take autonomously that decision, which allows full effectiveness in what they, by consequence, will undertake.*

3.3. An effective vocational decision: the necessary and sufficient condition

It is quite clear that society as such has to promote an optimal equilibrium between its societal needs and this psychological effectiveness of each of its members. That may motivate its government to take the counselling of its last year students in Secondary Education - it refers to the immediate future these youngsters themselves are planning - as seriously as possible. All kinds of measures can be developed in order to promote that, given the fact that this process of vocational decision psychologically turns out to be an individual process of matrix construction, selecting that ***environment*** that as a task effectively fits the ***person*** one is.

The crux is to meet, as effectively as possible, the 'different' needs these youngsters have at each of the up to five 'different' stages they are going through, in order to attain a firm decision with respect to their future on the immediate as well as on the long run. They vary from **(1)** a good explanation of the implications of the making of a vocational decision in the very first stage of **sensitisation**, to **(2)** the presentation of relevant information about occupations and their development within a modern society during the essential second stage of **horizon enlargement**, **(3)** the opportunity to test themselves with respect to their potentials, entering so their third stage of **self-concept clarification** with respect to each of the alternatives which seemed relevant at the end of the

preceding stage and last but not least - after **(4)** the making of their **decision** - the **(5)** Reality testing of what was decided during the very important 'last' stage of elaboration. Here the entrance exam for medicine and dentistry, by definition, has to fulfill, as the other ones in the faculties of engineering and in the artistic study domains of Higher Education outside universities already do, an essential condition for the promotion of 'deep level learning' or effective studying.

As a member of the Flemish community in Belgium as my country, I hope that my Minister of Education will realize himself in due time also, that the **timing** of this entrance selection is as important as the definition of its content. For that reason I propose to organize it's first session at Eastern and its second at the end of June. So students who will not be allowed to register in medicine or dentistry, can take *in due time* a new decision that offers them elsewhere that same 'effectiveness', which means 'personal happiness', they *really* are looking for.

References

Ausubel, D. (1963). The psychology of meaningful verbal learning. An introduction to school learning. San Francisco/London: Jossey Bass Inc.

CgAO (Contactgroep Academisch Onderwijs) (1993). Van secundair naar hoger onderwijs; is er een einde aan die jaarlijkse hecatombe? Leuven: Garant.

Heider, F. (1958). The psychology of interpersonal relations. New York: Wiley.

Janssen, P.J. (1996). Studeren, doceren en evalueren in studaxologisch perspectief. Leuven: Acco.

Janssen, P.J. (1996). Studaxology: the expertise students need to be effective in higher education. Higher Education, 31, 117-141.

Janssen, P.J. & De Neve, H. (1988). Studeren en doceren aan het hoger onderwijs; vakmanschap als leeropdracht. Leuven: Acco.

Marton F. & Säljö, R. (1976). On qualitative differences in learning. 1. Outcome and processes. 2. Out-come as a function of a learner's conception of the task. British Journal of Educational Psychology. 48, 4-11 and 115-127.

Minnaert A. & Janssen, P.J. (1992). Success and progress in Higher Education: a structural model of studying. British Journal of Educational Psychology, 62, 184-192.

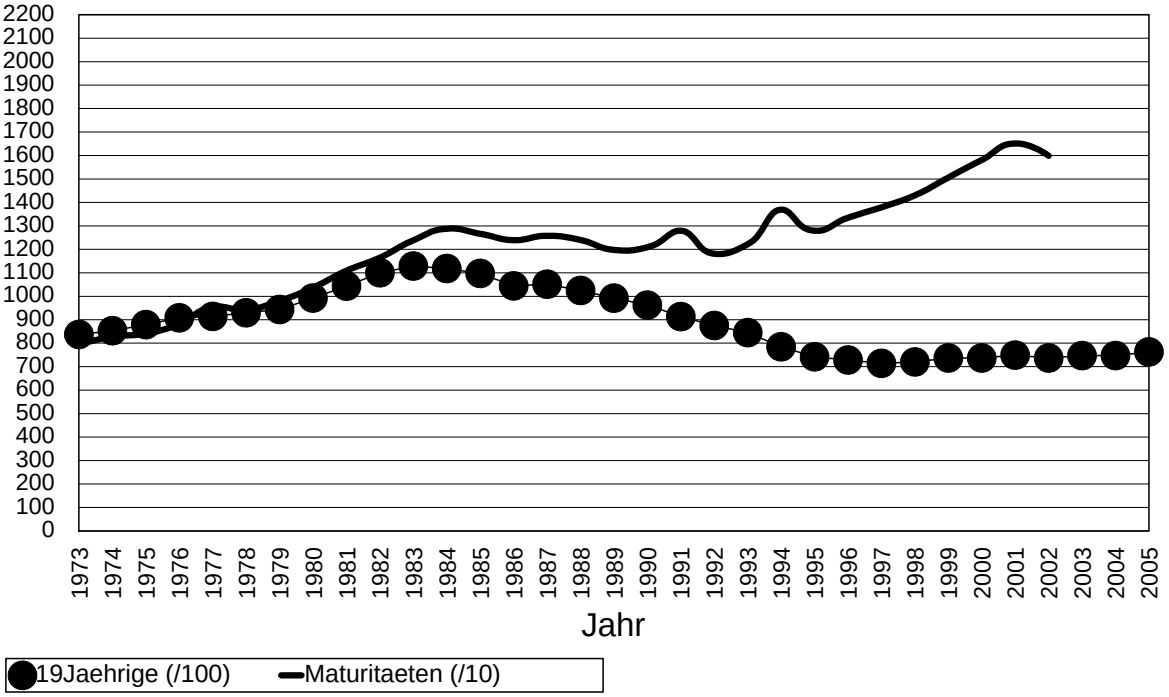
Pask. G. (1976). Conversational techniques in the study and poractice of education. British Journal of Educational Psychology, 46, 128-148.

Der Eignungstest für das Medizinstudium in der Schweiz als Instrument für die Beschränkung der Studienzulassung

K-D. Hänsgen

Wenn man die in diesem Band vorgestellten internationalen Tendenzen realistisch betrachtet, ist eine „Selbstlösung“ des Problems überschrittener Studienplatzkapazitäten im Fach Medizin in der Schweiz ziemlich unwahrscheinlich. In allen Industrieländern beobachtet man den verstärkten Wunsch nach besserer Ausbildung und ein Anwachsen der Zahl von Personen mit Zugangsberechtigung für ein Hochschulstudium. Dies gilt besonders für das Medizinstudium, weil der Arztberuf aus verschiedenen Gründen erstrebenswert bleibt. Deshalb haben die allermeisten Länder auf einen Numerus Clausus zurückgreifen müssen, um die Studienzulassung den Hochschulkapazitäten anzupassen.

Wie sind die Prognosen? In der folgenden Abbildung werden die Einschreibungen für das Fach Medizin den Maturitäten (geteilt durch 10) und der jeweils 19jährigen Wohnbevölkerung pro Jahr (geteilt durch 100) gegenübergestellt:



Die Maturitätszahlen ab 1996 sind der Schätzung des CSDOC (Schweizerische Dokumentationsstelle für Schul- und Bildungsfragen) entnommen. Trotz früher sinkender bzw. zukünftig gleichbleibender Bevölkerungszahlen nimmt die Zahl der Maturitäten beständig zu. Deshalb muss auch von einem bleibenden Druck auf die Hochschulen ausgegangen werden. Im Medizinstudium steigt die Zahl der Anfänger, einem eigenen Trend folgend, noch zusätzlich an.

Die Akzeptanz eines Numerus Clausus (NC) in der Schweiz wird um so höher sein, je fairer und wissenschaftlich abgesicherter das Kriterium für die „Regulierung" der Zulassung sein wird. Man muss sich allerdings entscheiden: Führt man einen Numerus Clausus ein, bedeutet dies definitiv Auswahl vor dem Studium. Es wäre die politische Entscheidung pro NC letztendlich nicht notwendig, wenn dann praktisch alle Studierwilligen dennoch aufgenommen werden sollen. Das Referendum im Kanton Bern im Jahre 1995 hat deutlich gezeigt, dass etwa 2/3 des Stimmvolkes für die Schaffung der gesetzlichen Grundlagen eines NC ist.

Nach welchen Kriterien kann die Zuteilung der Studienplätze erfolgen? Eine Selektion vor Studienbeginn erfordert nachvollziehbare, objektive und gerechte Kriterien, die auch einer juristischen Anfechtung standhalten müssen. Die folgende Tabelle fasst Vor- und Nachteile von vier möglichen Kriterien einmal zusammen:

	Maturität /Notendurchschnitt	**Eignungsgespräch**	**Sozialpraktikum**	**Eignungstest**
Gleichbehandlung/ Vergleichbarkeit	nicht erfüllt Benotung unterschiedlich streng mit unterschiedlichen Massstäben in den Kantonen und Schulen	Vergleich zwischen verschiedenen Gesprächsgruppen erfordert Standardisierung und Training der Beurteiler	vergleichbare Anforderungen und Betreuungsqualität nötig	durch Standardisierung der Durchführung voll gegeben
Zulassung entsprechend Kapazität möglich?	Ja	ja	Übereinstimmung wäre Zufall	ja
zusätzlicher Aufwand **Probleme**	keiner Problem: Rückwirkungen auf die Benotungsprinzipien der Schulen möglich	1 Termin pro KandidatIn, Training Urteiler, 3 Beurteiler 30 min mit Auswertung (1800*3*0,5 h) = 2530 Arbeitstage = 10 Arbeitsjahre (unrealistisch)	Zeitverlust für Kandidaten von einem Jahr möglich qualifizierte Betreuungskapazität zu gewinnen geeignete + bereite Einrichtungen suchen	1 Termin, Testleiter Auswertung erfolgt zentral
Begründung (Validitätsstudien)	entfällt wegen Benotungsunterschieden	weniger gute Prognose Studienerfolg als Test	„Abschreckung" - fehlende Erfahrung bezüglich Wirksamkeit	am besten Anwendbarkeit in der Schweiz ist geprüft
Wiederholbarkeit	nein	ja	nicht erforderlich	ja

Die Verwendung von **Maturitätsnoten** als Zugangskriterium hätte vor allem zwei sehr gravierende Nachteile: Zum einen würden die Unterschiede innerhalb und zwischen den Kantonen oder Sprachgruppen kaum gerecht auszugleichen sein. Diese zeigen sich beispielsweise in unterschiedlichen Maturitätsquoten (Anteil Maturitäten an den Schulabschlüssen eines Jahrganges), unterschiedlichen Benotungsmassstäben und einer verschieden gehandhabten Benotungsstrenge. Zum anderen würde es Rückwirkungen auf das Benotungssystem selbst geben, indem sich die Lehrenden natürlich der Funktion ihrer Noten für die Studienzulassung bewusst sind und dies kann dann zu verändertem Benoten führen.

Eignungsgespräche für alle Kandidatinnen und Kandidaten eines Jahrganges in der Schweiz führen sich durch einen jährlich notwendigen Aufwand, der ca. 10 Arbeitsjahren entspricht, von selbst ad absurdum. Sie bleiben in der Diskussion immer wieder „begehrt", weil man sich davon eine Überprüfung der Motivation und sozialen Kompetenz erwartet und man mehr den Aspekt der Berufseignung einbeziehen möchte. Aufgrund der insgesamt geringeren prognostischen Validität von Eignungsgesprächen für Studienerfolg (vgl. 18. Arbeitsbericht des ITB, Trost 1995) wäre Vorsicht geboten, die Zulassung in der Medizin **generell** davon abhängig zu machen. Das mag daran liegen, dass sich diese sozialen Kompetenzen während der Ausbildung erst entwickeln müssen. Was am ehesten möglich sein könnte, wäre eine Differenzierung im Bereich der Nicht-Eignung. Man sollte deshalb darüber nachdenken, sie für einen Teil der Bewerberinnen und Bewerber einzusetzen. Wir kommen unten darauf zurück

Der wichtigste Einwand gegen **Praktika** ergibt sich aus dem hohen Aufwand für die betreuenden Einrichtungen. Der Versuch im Kanton Zürich zeigt, dass diese zusätzlichen Belastungen für ein um Effektivität ringendes Gesundheitswesen nicht ohne weiteres zu erbringen sind und die Bereitstellung von genügend Plätzen nicht realisierbar ist. Der „eigentliche" Nachteil von Praktika besteht allerdings in der Unmöglichkeit, die Zahl der Zulassungen wirklich entsprechend der Kapazitäten zu regulieren. Sie können „zu dissuasiv" sein oder es begehren nach erfolgreich absolviertem Praktikum weiter zu viele Bewerber Eingang in die Universität. Hier würde der Zufall bestimmend sein.

Der Test als „Probestudium"

Betrachtet man die gesetzlichen Grundlagen in den Kantonen, ist die **Eignung** zum Studium das einzige praktikable wie akzeptable Kriterium für eine Zulassung. Dadurch wird gewährleistet, dass die Zugelassenen das Studium auch mit grosser Wahrscheinlichkeit beenden. Eine Beschränkung ohne Berücksichtigung der Eignung birgt das Risiko, dass weiter eine zu hohe Rate von Studienabbrüchen auftritt (weil die Zugelassenen den Anforderungen aus Lei-

stungsgründen nicht genügen) und die Kapazität der Hochschule sogar unterschritten werden könnte.

Scheiden Maturitätsnote und Eignungsgespräch als generelle Lösung zur Erfassung der Studieneignung aus, verbleiben nur Tests als sinnvolle Methoden, um diese Eignung zu erfassen. Psychologische Tests sind wissenschaftliche Prüfverfahren zur Untersuchung von Fähigkeiten oder Verhaltensweisen. Mit ihrer Hilfe werden Aussagen über spezifische Eigenschaften der Testpersonen gemacht, die eine Basis für Entscheidungen mit teilweise grosser Tragweite für diese Personen bilden können. Solche Aussagen müssen deshalb stichhaltig sein und bedürfen einer genau definierten wissenschaftlichen Überprüfung, bevor sie etwa zur Entscheidung über eine Studienzulassung oder die Eignung für eine bestimmte Tätigkeit angewendet werden können.

Ein psychologischer Test, der den wissenschaftlichen Anforderungen genügen will, muss mehrere Gütekriterien erfüllen:

An den Test wird der Anspruch gestellt, von der Person des Prüfers unabhängige Resultate zu erbringen, das heisst er muss objektiv sein. Die Chancengleichheit muss durch standardisierte Bedingungen bei der Durchführung und der Auswertung der Ergebnisse gewährleistet sein.

Bei einem Test, dessen Ergebnisse eine Aussage über einen zukünftigen Studienerfolg erlauben sollen, müssen die Testergebnisse auch tatsächlich mit diesem Studienerfolg in Beziehung stehen. Diese Überprüfung wird in der Psychologie als Validitätskontrolle eines Tests bezeichnet.

Der Test muss aber auch zuverlässig sein, das heisst die einzelnen Aufgaben je Untertest müssen dasselbe Merkmal, die einzelnen Untertests zusammen dieselbe Fähigkeit, messen.

Ein **Intelligenztest** hat den Anspruch, Fähigkeiten weitgehend allgemeingültig zu erfassen, ohne sich auf eine bestimmte Tätigkeit zu beziehen. Er prüft zum Beispiel, ob die Testperson ein ihrem Alter entsprechendes Wissen und eine entsprechende Handlungsfähigkeit aufweist. Mit diesen Resultaten kann beispielsweise die Schulreife überprüft werden. Der Intelligenztest misst deshalb ein breites, globales Wissen oder sehr allgemeine Fähigkeiten.

Demgegenüber handelt es sich bei einem **Eignungstest** um ein Verfahren, welches bezüglich einer konkreten Anforderung, einer speziellen Tätigkeit, verlässliche Resultate liefert. Auf deren Basis soll eine Vorhersage über die Bewährung in dieser Tätigkeit getroffen werden. Der Eignungstest überprüft beispielsweise, ob bestimmte zukünftige Belastungsereignisse in einer Tätigkeit gemeistert werden können.

Die Entwicklung des Tests für medizinische Studiengänge (TMS) in Deutschland, der ein solcher auf das Medizinstudium zugeschnittener Eignungstest ist, erfolgte mit dem Ziel, eine Prognose machen zu können, mit welcher Wahrscheinlichkeit der Testteilnehmer oder die Testteilnehmerin die Anforderungen des medizinischen Grundstudiums meistern wird.

Die Inhalte des Tests sollten dabei in einem sehr engen Zusammenhang zum Medizinstudium stehen. Deshalb handelt es sich bei den Testfragen um Inhalte aus den verschiedenen Anforderungen im medizinischen Grundstudium. Man hat sich auf den Leistungsaspekt beschränkt, weil es als nicht praktikabel erachtet worden ist, soziale und kommunikative Fähigkeiten zu berücksichtigen. Dies ergibt sich nicht nur aus den Schwierigkeiten, sie zuverlässig zu erfassen. Standardisierte Eignungsgespräche sind vor allem wegen fehlender Ressourcen nicht realisierbar. Auch die Dynamik der Entwicklung dieser Fähigkeiten im Alter zwischen Maturität und Studienabschluss und die damit verbundene geringe Prognosekraft solcher Aussagen liess davon Abstand nehmen.

Der geplante Eignungstest besteht aus neun Untertests, das heisst neun Gruppen von Testaufgaben desselben Typs, welche die Testpersonen am Tag der Testdurchführung in ca. 5 Stunden zu absolvieren haben. Fünf dieser Aufgabengruppen werden vormittags und vier nachmittags bearbeitet. Dazwischen ist eine einstündige Mittagspause vorgesehen.

Jeder Untertest beginnt mit einem kurzen Hinweis, in dem erklärt ist, was mit den jeweils folgenden Aufgaben geprüft wird. Alle Aufgaben eines Untertests sind nach dem gleichen Prinzip konstruiert. Mit Ausnahme des Konzentrationstests sind zu jedem Problem fünf mögliche Antworten vorgegeben; aber nur eine ist im Sinne der Aufgabenstellung richtig. Für das Bearbeiten jedes Untertests steht den Testpersonen eine begrenzte Zeit zur Verfügung.

Struktur und Ablauf des Tests für Medizinische Studiengänge

Quelle: Trost, Günter (Hrsg.) (1994): Test für Medizinische Studiengänge (TMS). Studien zur Evaluation. 18. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Bezeichnung der Untertests	geprüfte Fähigkeiten	Zahl der Aufgaben *	Bearbeitungszeit (in Minuten)
Muster zuordnen	differenzierte visuelle Wahr-nehmung	24 (20)	22
Medizinisch-naturwissenschaftliches Grundverständnis	Verständnis für medizinisch-natur-wissenschaftliches 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	1200 Zeichen	8
M I T T A G S P A U S E 60 Minuten			
Lernphase zu den Gedächtnistests:			
Figuren lernen			4
Fakten lernen			6
Textverständnis	Verständnis und Interpretation medizinischer und naturwissenschaftlicher Texte	24 (18)	60
Reproduktionsphase:			
Figuren lernen	Behalten von figuralem Material	20 (20)	5
Fakten lernen	Behalten von verbalem Material	20 (20)	7
Diagramme und Tabellen	Interpretation von Diagrammen und Tabellen	24 (20)	60
Gesamttest		204 (178)	5 Std. 7 Min.
Gesamtdauer der Testabnahme		etwa von 8.45 bis 16.00 Uhr	

* in Klammer sind die gewerteten Aufgaben angegeben; bei den übrigen Aufgaben handelt es sich um Einstreuaufgaben, die unter Ernstfallbedingungen für einen der folgenden Tests überprüft werden.

Ist der Test spezifisch genug für die Schweiz?

Die Adaptierung des Tests aus Deutschland bietet zunächst eine Gewähr, dass nicht bei Null begonnen werden muss und mögliche Irrwege beschritten werden, sondern dass sofort die für eine Vorhersage des Studienerfolges geeigneten Tests zur Verfügung stehen. Andererseits bedeutet dies nicht, dass schweizerische Besonderheiten keine Berücksichtigung finden. Die Dringlichkeit einer Lösung für die Schweiz machte es erforderlich, dass sehr schnell ein guter Test bereitstand. Hier hält der deutsche Test vor allem im internationalen Vergleich stand. Eigenentwicklungen „von Null an“ würden mindestens 2 Jahre Vorlauf erfordern. Ohne Erprobung unter Ernstfallbedingungen würde man keine einzige Aufgabe in einem Test einsetzen können. Nach Auskunft der Entwickler erfüllen nur ca. 1/3 der „am grünen Tisch“ von den besten Experten erarbeiteten Aufgaben dann tatsächlich die genannten Kriterien einer Anwendbarkeit im Test. Die Schwierigkeit der Aufgaben lässt sich auch von den Experten nur grob schätzen. Erst eine Erprobung unter Ernstfallbedingungen - eingestreut in einen Originaltest - gewährleistet, dass die Leistungen der Bewerberinnen und Bewerber optimal und zuverlässig differenziert werden können.

Im Juni 1995 wurde in Freiburg eine empirische Untersuchung durchgeführt, bei der eine Vorform des Tests unter ernstfallnahen Bedingungen an einer kleinen Stichprobe erprobt wurde (Hofer, Ruefli, Hänsgen 1995). Zur Überprüfung der psychologischen Gütekriterien und der Fairness des Tests wurde im Collège Sainte-Croix in Freiburg (Schweiz) der Test mit einer Stichprobe von 54 deutschsprachigen und 126 französischsprachigen Gymnasiastinnen und Gymnasiasten absolviert.

Dabei ist zu beachten, dass es sich bei diesem Probelauf nicht um eine Bewerbungssituation im üblichen Sinne gehandelt hat. Den Teilnehmerinnen und Teilnehmern konnte als “Motivierung“ nur eine individuelle Leistungsrückmeldung angeboten werden.

Trotz dieser fehlenden Bewerbungsmotivation bei der Stichprobe erreichte der Test annähernd **gleiche Gütekriterien** der Zuverlässigkeit wie in Deutschland. Es wurden im Mittel etwa genau die Hälfte (48%) der gewerteten Aufgaben richtig gelöst und die Streuung der Werte befindet sich in einem für die **Leistungsdifferenzierung optimalen** Bereich.

Beide **Sprachgruppen** erreichten dabei gleich gute Z**uverlässigkeitswerte** und es kann von einer hohen Äquivalenz beider Formen ausgegangen werden.

Im Hinblick auf die Prognosekraft des Testergebnisses für den zukünftigen Prüfungserfolg in den ärztlichen Vorprüfungen zeigte es sich, dass bei der französischsprachigen Gruppe die Test-Mittelwerte für die Absolventinnen und Ab-

solventen des Maturitätstypus D signifikant niedriger liegen als die für die Typen A, B und C. Das würde mit dem Prüfungsverhalten im Medizinstudium in Bern übereinstimmen, wo zwischen 44% und 50% der Absolventinnen und Absolventen von Maturitätstyp A bis C das erste Propädeutikum im ersten Anlauf nicht bestehen, bei Typ D sind es dagegen 77% (Hofer 1992).

Insgesamt belegen die Ergebnisse des Probelaufs, dass der **Test auch in der Schweiz als faires und zuverlässiges Zulassungskriterium** verwendbar wäre. Besorgnis bezüglich Leistungsunterschieden zwischen den Geschlechtern, den Sprachgruppen oder gar im Verhältnis zu Deutschland als dem Land, in dem der Test entwickelt worden ist, lassen sich nicht bestätigen.

Gewährleistet der Test die Gleichbehandlung (Chancengleichheit)?

Gegenüber allen anderen Kriterien bietet der Test vor allem die Voraussetzungen, eine strikte Gleichbehandlung aller Bewerberinnen und Bewerber zu realisieren. Das ist eine für die Schweiz besonders wichtige Voraussetzung.

Der Test wird zur gleichen Zeit an allen Testorten unter vergleichbaren Bedingungen durchgeführt. Alle Kandidatinnen und Kandidaten haben anhand der veröffentlichten Informationsbroschüre über den Test die Chance, sich ausreichend auf die Situation vorzubereiten. Die Durchführung ist standardisiert, Raumanforderungen, zeitliche Abläufe und die mündliche Instruktion der Teilnehmerinnen und Teilnehmer sind genau vorgegeben. Das mit der Testdurchführung beauftragte Personal wird vorher ausreichend geschult. Die Auswertung ist objektiv und es gibt keine Zweifel darüber, was richtig oder falsch ist.

In die Beurteilung gehen keine subjektiven Massstäbe ein, was sich beispielsweise in Gesprächssituationen nicht oder nur sehr schwer vermeiden liesse.

Bei Verwendung eines Tests der hier vorgeschlagenen Art wird auch das Problem einer möglichen Benachteiligung von Jugendlichen aus „nichtprivilegierten“ Schichten relativiert. Die soziale Schichtzugehörigkeit wird in Deutschland beim TMS seit Jahren laufend mit erfasst. Es kann nachgewiesen werden, dass das Verhältnis der Zugelassenen pro Schicht immer ziemlich genau der Bewerbersituation entspricht. Ein Grund dafür kann sein, dass mögliche Benachteiligungen in der Schule ausgeglichen werden können, weil eben nicht nur die Menge erworbenen Wissens erfasst wird. Die Schülerinnen und Schüler haben faktisch eine „neue Chance“.

Gilt die Chancengleichheit auch für die drei Sprachgruppen?

Chancengleichheit muss besonders für die Sprachgruppen gelten. Damit verbunden ist natürlich auch die Frage der Gleichheit der drei Sprachversionen im Schwierigkeitsgrad. Es wurde ein den international üblichen Standards entsprechendes Verfahren für die Übersetzungen angewendet, welches die folgenden Schritte beinhaltet:

- Erstübersetzung durch einen Fachübersetzer aus dem Wissenschaftsgebiet, dessen Muttersprache die Zielsprache ist;
- Begutachtung dieser Übersetzung durch Fachexperten (die zweisprachig sind);
- Rückübersetzung durch einen vom Erstübersetzer unabhängigen Zweitübersetzer in die deutsche Sprache und Vergleich der Rückübersetzung mit dem Ursprungstext;
- Konferenz aller Beteiligten zur Diskussion der Diskrepanzen und zum Finden der endgültigen Formulierung entsprechender Textabschnitte.

Als Ergebnis liegen drei äquivalente Sprachformen nunmehr vor, die vom wissenschaftlichen Standpunkt her gesehen eine hohe Chancengleichheit garantieren. Der Test könnte in deutscher, französischer oder italienischer Sprache absolviert werden und die Ergebnisse bleiben äquivalent.

Werden Frauen durch den Test benachteiligt?

Zu Recht wird gefordert, dass Männer und Frauen die gleichen Zugangschancen zur Hochschule finden müssen und dass die erreichten Fortschritte durch die Anwendung des Tests nicht gefährdet werden dürfen. Von verschiedenen Seiten wurde der Verdacht geäussert, dass der Test Frauen benachteiligen könne. Das ist bei genauer Betrachtung aber nicht ganz richtig. Vor allem müssen bei der Zulassung keine Nachteile für Frauen entstehen:

Bei der Bewertung des Tests wird jeweils über die Ergebnisse mit dem TMS in Deutschland gesprochen. Es ist dort Realität, dass Frauen im statistischen Notendurchschnitt die Medizinprüfungen schlechter abschliessen als Männer. Das wurde über lange Jahre in zahlreichen Analysen bestätigt. Ein zweites Faktum ist aber, dass Frauen auch im Test etwa um 2 Punkte (bei einem Mittelwert von 100 und einer Standardabweichung von 10) schlechtere Werte erreichen als Männer. Relativ gesehen ist der Geschlechterunterschied bei den Prüfungsnoten übrigens grösser als im Test. Wenn der Test ein „Probestudium“ sein soll, welches die Studienleistungen tatsächlich vorhersagt, dann konnte man dies er-

warten - ja, dann muss dies sogar so sein, wenn der Test die Realität unverzerrt widerspiegeln soll.

Es kommt aber darauf an, was man aus den Fakten macht und wie man mit ihnen umgeht. Die Feststellung dieses Unterschiedes bedeutet nicht automatisch, diesen Unterschied so auch in ein Zulassungskriterium zu übernehmen. Wenn der politische Entscheid dazu gefasst wird, kann ein statistisches Korrekturverfahren angewendet werden, welches den Mittelwert-Unterschied zwischen den Geschlechtern ausgleicht. Frauen hätten danach im Mittel die gleichen Testwerte wie Männer. Die Zulassungsquoten würden genau den Bewerbungsquoten für beide Geschlechter entsprechen.

Die Anwendung dieses Verfahrens erfordert allerdings neben der politischen Entscheidung dafür auch den Nachweis, dass die genannten Unterschiede wirklich für die Schweiz zutreffen. Die aus den deutschen Ergebnissen abgeleitete Hypothese, dass **Frauen tendenziell schlechtere** Leistungen erreichen als Männer, konnte beim Probelauf des Tests in der Schweiz **nicht bestätigt** werden. Bei der deutschsprachigen Gruppe erzielten die Frauen sogar bessere Ergebnisse als die Männer. Bei der französischsprachigen Gruppe war kein signifikanter Unterschied in den Ergebnissen beider Geschlechter vorhanden. Deshalb sollten die Ergebnisse der Anwendung unter Realbedingungen abgewartet werden.

Sind Trainingskurse notwendig?

Ein spezifischer Vorzug des deutschen Tests ist, dass sich ein Training zusätzlich zur empfohlenen Vorbereitung ganz eindeutig nicht lohnt. Dies haben mehrere Untersuchungen des Instituts für Test- und Begabungsforschung in Bonn sowie eines unabhängigen Evaluationsgremiums wissenschaftlich nachweisen können. Die empfohlene Vorbereitung beschränkt sich auf das Durcharbeiten einer Informationsbroschüre über den Test und die Bearbeitung einer veröffentlichten Originalversion. Zusätzliche Trainingskurse oder Trainingsbücher kommerzieller Anbieter bringen keine nennenswerte zusätzliche Leistungssteigerung (siehe Beitrag von Hofer und Hänsgen in diesem Heft).

Ein Wort zu den Kosten der Testanwendung

Das Kostenargument gegen eine „Zulassungsbürokratie“ wäre kurzsichtig - es sind vergleichsweise einfache Strukturen in der Schweiz notwendig. Bei einem Etat von 850’000,- Fr. und geschätzten Kosten von 50’000,- Fr. pro Studienplatz im Medizin-Grundstudium würde dies den Kosten von 17 Studienplätzen entsprechen. Nur etwa 10% der durch eine anfängliche gesamtschweizerische Regelung der Zulassung insgesamt eingesparten Mittel würden für das Verfah-

ren selbst benötigt. Dieser „Wirkungsgrad“ von 90% wäre nicht der schlechteste. Er erhöht sich weiter, wenn man die Mittel für die 60% Studienabbrecher in den ersten beiden Jahren noch mit berücksichtigt. Ein beträchtlicher Teil dieser Mittel würde ebenfalls nicht mehr notwendig sein.

Schlussfolgerungen und mögliche Entwicklungen für die Zukunft

Bei der Entscheidung, **aufgrund welcher Kriterien** ein Numerus Clausus in der Schweiz **realisiert** werden kann, wird ein **Eignungstest nach wie vor favorisiert** - wie beispielsweise in den USA und Kanada, Schweden, Tschechien, Finnland, Israel, Japan bzw. als standardisierter Wissenstest in 17 weiteren europäischen Ländern. Wegen der spezifischen Situation der Schweiz (kaum vergleichbare Maturitätsnoten zwischen den Kantonen resp. Sprachregionen, Unmöglichkeit von Eignungsgesprächen für alle Bewerberinnen und Bewerber allein aus Kostengründen, Realisierungsprobleme für voruniversitäre Praktika) bleibt ein solcher Eignungstest - zugeschnitten in seiner Anwendung auf die Bedingungen der Schweiz - das **fairste und wissenschaftlich gesichertste Kriterium**.

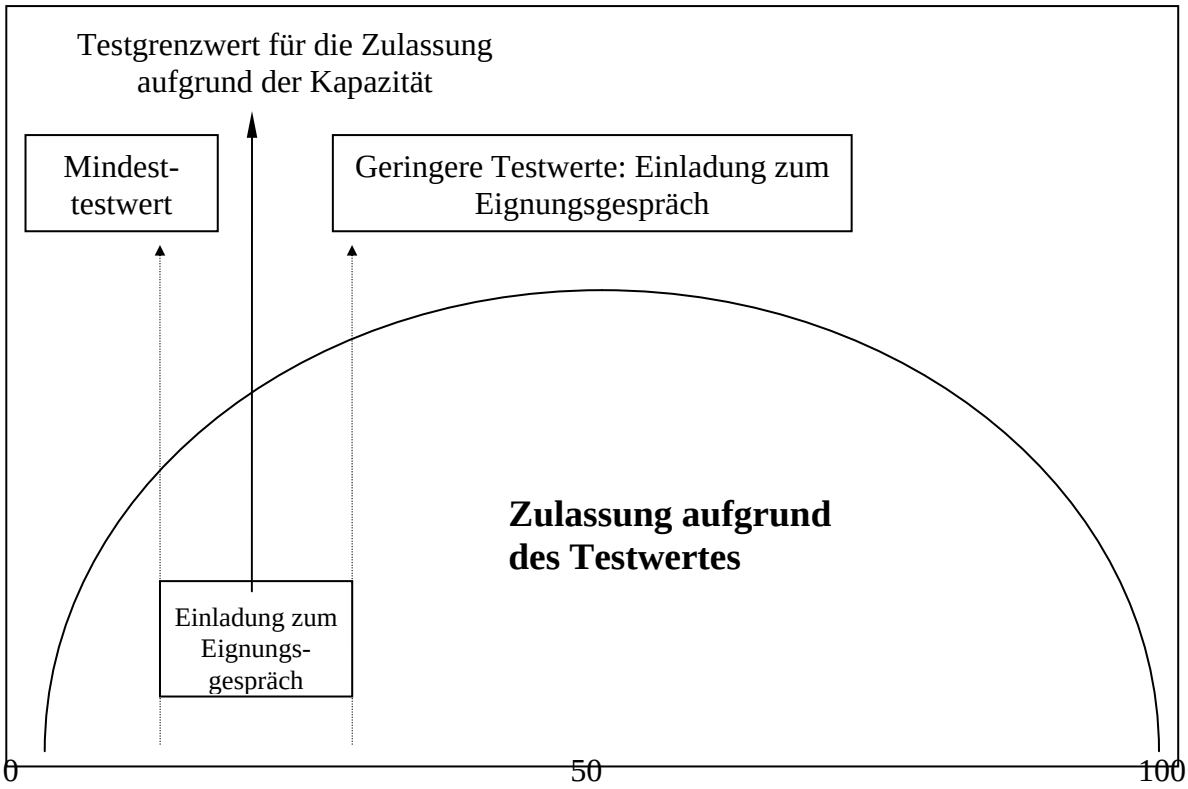
Der deutsche Test TMS als Grundlage des Schweizer Verfahrens wurde sehr aufwendig konstruiert und elaboriert, er schneidet im internationalen Vergleich bezüglich der Vorhersage von Studienerfolg mit am besten ab. Die **Nachnutzung** eines vorhandenen deutschen Tests in der Schweiz verursacht weniger als 1/10 der Entwicklungskosten in Deutschland - bei gleicher Qualität der Erfolgsvorhersage. Dies ist eindeutig ein Vorteil des hier in der Schweiz gewählten Konzeptes. Jede Testform muss jährlich erneuert werden, damit die Anforderungen garantiert bei Kandidatinnen und Kandidaten nicht bekannt sind.

Bei Notwendigkeit eines NC in der Schweiz wird längerfristig zu prüfen sein, (1) ob **alternative Testkonzepte bei gleicher Güte weniger aufwendig** entwickelt werden können und (2) **ob Ergänzungsmethoden sinnvoll sind.** Wenn die Studieneignung um weitere Merkmale der Berufseignung ergänzt werden soll, wären standardisierte Eignungsgespräche allenfalls für einen Teil der Bewerber eine praktikable Methode, um Motivation und soziale Kompetenz zu „prüfen“. Daran sind allerdings relativ hohe Anforderungen gebunden, etwa die Vergleichbarkeit von Ergebnissen zwischen verschiedenen Interviewergruppen und die nachvollziehbare wie wissenschaftlich gesicherte tatsächliche Regulierbarkeit der Zulassung (etwa durch einen Punktwert analog zum Test).

Es scheint heute aus ökonomischen Gründen unrealistisch, gesamtschweizerisch pro Jahr ca. 2000 Eignungsgespräche zu führen. Deshalb wäre die **Kombination eines obligatorischen Tests mit Zulassungsgesprächen** für eine begrenzte Zahl der Bewerber überdenkenswert. Auf diese Art könnte die grosse

Mehrzahl der Studienplätze nach der Testleistung vergeben werden. Dabei würde es sich um diejenigen handeln, deren Leistungen für das Bewältigen der Anforderungen eines Medizinstudiums am ehesten ausreichen.

Es würden diejenigen zum Gespräch eingeladen, die aufgrund des **Testergebnisses keinen Studienplatz** erhalten können. Entweder man bietet dies allen an oder man setzt auch hier ein Kriterium an, ob ein **Mindest-Testwert** erreicht wurde. Dies böte die Möglichkeit, **sehr gezielt einen Personenkreis** einzuladen, der Testwerte in einem definierten Bereich um den „Grenzwert" der pro Jahr möglichen Zulassungen erreicht hat. Die notwendige Kapazität für Eignungsgespräche pro Jahr liesse sich auf diese Weise sehr genau vorher festlegen. Das folgende Schema soll dies verdeutlichen:



Unter dem zum Eignungsgespräch eingeladenen Personenkreis werden die restlichen Studienplätze nach Massgabe des Zulassungsgesprächs an die Geeignetsten vergeben. Dabei können die Testleistungen in gewichteter Form noch mit berücksichtigt werden und es wäre dadurch ein Ausgleich schlechterer Testleistungen durch bessere soziale Kompetenzen möglich.

Die Eignungsgespräche müssten standardisiert sein. Die Themenbereiche der Fragen werden vorgegeben, 3 Beurteiler sollten an jedem Gespräch beteiligt sein und anschliessend ihren Eindruck in einem standardisierten Bewertungs-

bogen nach verschiedenen Kriterien einschätzen. Dieser führt ebenfalls zu einem Gesamt-Punktwert, der Personen über die verschiedenen Beurteilergruppen vergleichbar macht (als Summe der Punkte über die 3 Beurteiler). Systematische Beurteiler-Unterschiede zwischen den Gruppen könnten statistisch ausgeglichen werden.

Die Leistungen für die Erstellung des Konzeptes und der Unterlagen wären vom ZTD vermutlich nur mit geringen Mehrkosten zu erbringen, da zusätzliche Personalkosten nur in geringem Masse notwendig scheinen.

Dennoch plädieren wir dafür, zunächst den Eignungstest als Kriterium zu verwenden. Er ist wissenschaftlich überprüft, in der Schweiz anwendbar und liegt in drei äquivalenten Sprachformen vor. Das mit seiner Hilfe realisierbare Zuordnungsverfahren ist fair und - wie sich anlässlich dieser Tagung zeigt - international anerkannt. Die Suche nach Perspektiven und Erweiterungen, die zusätzliche Aktivitäten und Vorbereitungen erfordert, sollte nicht verhindern, dass heute etwas getan wird - weil es getan werden muss.

Literatur

Hänsgen, K.-D., Hofer, R., Ruefli, D. (Hrsg.) (1996) Eignungsdiagnostik und Medizinstudium. Tagungsband. Berichte des ZTD Band 2 Univ. Fribourg.

Hänsgen, K.-D., Hofer, R., Ruefli, D. (1996). Un test d'aptitudes aux études de médecine est-il faisable en Suisse? Bulletin des médecins suisses, 7, S. 267 - 274.

Hänsgen, K.-D., Hofer, R., Ruefli, D. (1995). Der Eignungstest für das Medizinstudium in der Schweiz. Schweizerische Ärztezeitung, 37, S. 1476 - 1496

Hofer, R., Ruefli, D., Hänsgen, K.-D.(1996). Der Eignungstest für das Medizinstudium in der Schweiz. Ein Probelauf. Berichte des ZTD Band 1 Univ. Fribourg

Trost, Günter (Hrsg.) (1978-1994): Test für Medizinische Studiengänge. 1. - 18. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Zentrum für Testentwicklung (1995). Il test attudinale per lo studio della medicina (Adattamento italiano). Göttingen: Hogrefe

Zentrum für Testentwicklung (1995). Le test d'aptitudes pour les études de médecine (Adaptation française). Göttingen: Hogrefe

Die Trainierbarkeit von Testleistungen im Zusammenhang mit einem Eignungstest für das Medizinstudium in der Schweiz

R. Hofer & K.-D. Hänsgen

Die mögliche Einführung eines Eignungstests als Auswahlverfahren zum Medizinstudium in der Schweiz rückt die Frage nach der Trainierbarkeit von Testleistungen ins Blickfeld der Öffentlichkeit. Die Testabsolvierung darf nicht zu unvertretbaren finanziellen oder zeitlichen Zusatzbelastungen für die Kandidatinnen und Kandidaten führen. Werden die Zulassungschancen durch einen zusätzlichen Besuch von Trainingskursen signifikant verbessert, würde weniger die Studieneignung die Zulassung bestimmen. Einen grösseren Einfluss hätte unter diesen Umständen die Bereitschaft und Fähigkeit, Ressourcen für ein Training aufzuwenden.

Beim verwendeten Testkonzept des deutschen Tests für Medizinische Studiengänge (TMS) wurde Wert darauf gelegt, die Fähigkeit zum Erarbeiten neuer Lösungen und Anforderungen zu prüfen und nicht „eingepauktes“ Wissen abzufragen. Alle zur Lösung einer Aufgabe notwendigen Wissenselemente werden in der Aufgabe selbst mitgeteilt. Da bestimmte Übungseffekte bei einzelnen Aufgaben bekannt sind und natürlich auch eine Vertrautheit mit der Art der Aufgabenstellung die Leistungen beeinflusst, erfolgt die Vorbereitung auf den Test mittels einer Broschüre, welche Beispielaufgaben und Übungshinweise enthält. Sie enthält nach Aussagen der Testentwickler die notwendigen Vorbereitungsmassnahmen - und nur diese. Zu fragen ist, ob diese „standardisierte“ Vorbereitung ausreichend ist und weitere Massnahmen keine zusätzlichen Effekte bringen. Gilt letzteres, kann und sollte die Auseinandersetzung mit Anbietern von Testtraining intensiv und offensiv erfolgen. Leider zeigt die Erfahrung aus Deutschland, dass die Argumentation dieser Ausbildungsanbieter manchmal stärker auf die Gewinnung von (zahlenden) Teilnehmern gerichtet ist als auf die Realität.

Es darf davon ausgegangen werden, dass nach der Einführung des Eignungstests für das Medizinstudium in der Schweiz auch hierzulande mit dem kommerziellen Angebot von zusätzlichen Trainingsbüchern und Trainingskursen zum Test zu rechnen ist. Für die Nutzung dieser Angebote wird voraussichtlich mit dem Argument geworben, dass durch das Testtraining die Leistung und mithin die Zulassungschancen zum Medizinstudium deutlich erhöht würden oder man den Test gar nur durch den Besuch dieser Kurse "bestehen“ könne. Diese Werbungen nutzen letztendlich die objektiv vorhandene Testangst der Kandidatinnen und Kandidaten, die in der für die Schweiz neuen Situation verständlich und in bestimmtem Grade auch unvermeidlich scheint. In einer Zeit der Ersteinführung eines Tests in der Schweiz treffen Test-Unerfahrene dann

auf langjährig in Deutschland tätige Test-Trainer, deren Werbeaussagen in langjähriger Praxis erprobt worden sind.

Dass eine stärkere Motivation die Chancen einer Zulassung erhöht, wenn sie durch einen höheren Vorbereitungsaufwand auf den Test unter Beweis gestellt wurde, müsste für sich genommen dabei nicht einmal schlecht sein. Sollte aber die Zahlungskraft die Intensität dieser Vorbereitung beeinflussen, dann könnten Zahlungskräftigere einen ungerechtfertigten Vorteil gegenüber anderen erlangen. Dies wäre definitiv nicht gewollt. Auch in Deutschland wurde die Frage nach dem Trainingseinfluss deshalb früh gestellt - verbunden mit der Forderung nach ihrer wissenschaftlichen Überprüfung (zusammenfassend siehe Deter 1982, S. 12 f.).

Erste Vermutungen, dass die Aussagekraft von Intelligenztests durch Training beeinflusst werden könnte, findet man schon bei Thorndike (1919, zitiert in Deter 1982, S. 14): „In proportion as such tests are used for promotion in schools, entrance examinations for colleges or professional schools, employment examinations in business and industry, or any purpose where a high score is to the advantage of the person tested, the danger from deliberate coaching become very grave."

Seither wurde immer dann der Frage der Trainierbarkeit von Testleistungen nachgegangen, wenn sich bestimmte Kreise der Gesellschaft einem Test zu unterziehen hatten und wenn diese Resultate zu einer Selektion herangezogen wurden. So geschehen beispielsweise in den USA, als in den zwanziger Jahren der "Army Alpha", der erste Intelligenztest für Gruppenuntersuchungen, für den zivilen Gebrauch freigegeben wurde oder Ende der vierziger Jahre in Grossbritannien, als die "Even-plus"-Prüfung für die Aufnahme in die britischen Gymnasien eingeführt wurde (s. Deter, 1982). Gleiches geschah, als Ende der 50er Jahre in den USA die Bedeutung der "Scholastic Aptitude Tests" (SAT) für die Zulassung zu Colleges zunahm oder Ende der 70er Jahre in Deutschland, als der Test für medizinische Studiengänge (TMS) seinen ersten Probelauf vor sich hatte.

Der TMS besteht aus neun Untertests, das heisst neun Gruppen von Testaufgaben jeweils desselben Typs. Die Inhalte des Tests stehen in einem sehr engen Zusammenhang zum Medizinstudium (siehe Trost et al. 1976-1996). Geprüft werden Fähigkeiten der Wiedererkennung (Untertest "Muster zuordnen"), das Verstehen von Fragen ("Medizinisch-naturwissenschaftliches Grundverständnis"), das räumliche Vorstellungsvermögen ("Schlauchfiguren"), der Umgang mit Zahlen, Grössen, Einheiten und Formeln ("Quantitative und formale Probleme"), das konzentrierte Arbeiten ("Konzentriertes und sorgfältiges Arbeiten"), das Einprägen ("Figuren lernen", "Fakten lernen"), die Aufnahme und

Verarbeitung von Textmaterial ("Textverständnis") und die Analyse von Diagrammen und Tabellen ("Diagramme und Tabellen").

Es ist klar, dass Vorab-Beschäftigung mit dem Test Vorteile bringt. Allein durch vorherige Vertrautheit mit den Instruktionen gewinnt man zusätzliche Lösungszeit im „Ernstfall". Bei einigen Aufgaben können bestimmte Hinweise die Lösungsfindung vereinfachen. Für einzelne Tests, beispielsweise den Konzentrationstest, sind durch mehrmaliges Üben die Leistung zu verbessern.

In Deutschland wird deshalb kostenlos eine Test-Broschüre an alle Betroffenen abgegeben, wo standardisierte, zumutbare und nach Meinung der Testentwickler ausreichende Vorbereitungsmassnahmen zusammengestellt sind. Die standardisierte Vorbereitung mit dieser Broschüre beinhaltet:

- Kenntnis über den Ablauf des Tests (Pausen, Verpflegung, notwendige Dinge);
- Kenntnis der einzelnen Aufgabengruppen anhand von Beispielaufgaben und Hinweisen zur Lösung - mit der Möglichkeit, einzelne Musteraufgaben selber zu lösen; zu jedem Untertest des TMS sind die Originalinstruktion sowie mehrere Aufgabenbeispiele samt Lösungen und Erklärungen abgedruckt.
- genaue Erklärung, wie ein Test insgesamt funktioniert, wie die Testwerte gebildet werden und welche Strategien bei der Bearbeitung sinnvoll sind;
- der Hinweis, dass man den Konzentrationstest mehrfach selbst üben soll;
- der Hinweis, dass man eine veröffentlichte Originalversion unter realistischen Zeitbedingungen einmal selbst bearbeiten soll; Zu diesem Zwecke sind zwei Originalversionen veröffentlicht worden.

Diese Hinweise sind im Ergebnis einer laufenden Evaluation der Testanwendung in Deutschland laufend überarbeitet und aktualisiert worden.

Die von dritter Seite angebotenen Testtrainings werden sehr unterschiedlich durchgeführt und machen deshalb einen Vergleich der Effekte schwierig (s. Deter 1982, S. 45 ff.). Die Kurszeiten liegen zwischen nicht einmal einstündigen Sitzungen und einem ganzen Jahr. Inhaltlich bewegen sich die Angebote zwischen simplen Itemübungen und gezielten, intensiv durchgeführten Kursen, in denen Lösungsstrategien abgegeben und geübt werden, um die Teilnehmenden möglichst optimal vorzubereiten. Dazu kommt, dass nebst den Trainingskursen auch die Evaluationsstudien unterschiedlich durchgeführt werden (mit/ohne Kontrollgruppe, mit/ohne Vortestung, Testpersonen aus unterschiedlichen Populationen, etc.). Mit Ausnahme der Studie von Deter (1980, 1982)

fehlen auch die Angaben der Vorbereitungsmöglichkeiten für die Kontrollgruppe, die am Testtraining nicht teilgenommen haben.

Die Gründe, die zum Besuch eines Trainingskurses führen, liegen darin, dass laut Selbsteinschätzung sich die Mehrzahl der Kursbesucher und -besucherinnen nicht in der Lage sehen, eigenständig und ohne Anleitung konzentriert über mehrere Stunden das Trainingsmaterial zu bearbeiten (s. Mispelkamp 1987). Dazu hält Allalouf (1996) fest, dass es den Trainingsbesuchern und -besucherinnen vor allem darum geht, sich mit den Instruktionen, den Items, den Zeitgrenzen und den Antwortblättern des Tests vertraut zu machen und zusätzlich soviel wie möglich Unterlagen zu erhalten, was den Test anbelangt. Die Teilnehmerinnen und Teilnehmer erhoffen sich, ihre "test-wiseness" zu erhöhen, das heisst die Anwendung optimaler Strategien oder Techniken der Testbearbeitung mit dem Ziel kennenzulernen, unabhängig vom Inhalt der jeweiligen Testaufgaben den Testwert zu erhöhen. Bei der Auswertung von sieben Beobachtungsprotokollen, in denen die überwiegende Zahl der in Deutschland angebotenen Trainingskurse beschrieben sind (Mispelkamp 1987), wurden die angegebenen Strategien der Trainingsinstitute untersucht. Diese Strategien (Zeiteinteilungs-, Fehlervermeidungs- und Ratenstrategien sowie die Strategie für deduktives Schliessen) gehen kaum über die Bearbeitungshinweise und Ratschläge, welche die Zentralstelle für die Vergabe von Studienplätzen in der Test-Broschüre (ZVS 1995) angibt, hinaus. Es scheint aber hilfreich zu sein, diese Strategien vor der Testteilnahme kennenzulernen (s. Deter 1982, S. 24 ff.). Ganz klar zeigt es sich, dass den Bewerbern und Bewerberinnen, die früher bereits Erfahrungen in der Bearbeitung von Multiple-choice-Aufgaben sammeln konnten, den Einstieg in die Testsituation leichter gelingt als jenen, die einem standardisierten Testverfahren zum ersten Mal gegenüberstehen (Trost et al. 1977-1996).

Die Überprüfung der Effekte der einzelnen Vorbereitungsmittel gehören bei jedem Testtermin zu den Analysen der Zusammenhänge zwischen der Höhe der Testleistung und verschiedenen Merkmalen der Teilnehmer und Teilnehmerinnen (Trost et al. 1977-1996). Anhand der Antworten auf die Frage "Haben Sie schon einmal Testaufgaben, wie sie im Medizinertest vorkommen, bearbeitet" lassen sich die Bewerber und die Bewerberinnen in fünf Gruppen einteilen: (A) die sich noch nie mit solchen Testaufgaben befasst haben, (B) die sich nur mit Hilfe der Broschüre vorbereitet haben, (C) die sich mit Hilfe der Test-Broschüre und einer veröffentlichen Originalversion vorbereitet haben, (D) die sich mit Hilfe der Broschüre, einer veröffentlichen Originalversion und einem Trainingsbuch vorbereitet haben und (E) die sich unter anderem mit Hilfe eines Trainingskurses oder eines -seminars vorbereitet haben.

Der Einfluss der Vorbereitungsmittel auf die Testergebnisse im Durchschnitt der 6 Testtermine aus den Jahren 1982-84.

Gruppe	Vorbereitung	Durchschnitt*
(A)	keine	94,7
(B)	Broschüre	99,4
(C)	Broschüre + öffentliche Originalversion	104,4
(D)	Broschüre + öffentliche Originalversion + Trainingsbuch	105,8
(E)	Broschüre + öffentliche Originalversion + Trainingskurs	104,4

*Mittelwert 100, Standardabweichung 10

Fasst man die Ergebnisse der sechs Testtermine aus den Jahren 1982 bis 1984 zusammen, so zeigt es sich, dass diejenigen, die sich der Gruppe A zuordneten, im Mittel mit Abstand am wenigsten Punkte erzielten, nämlich 94,7 bei einem Mittelwert von 100 und einer Standardabweichung von 10 Punkten. Im Mittel am meisten Punkte (105,8) erzielte die Gruppe Testpersonen, die sich nach eigenen Angaben nach den Kriterien der Gruppe D vorbereitet haben. Während die Gruppen C und E mit ihren Mittelwerten von je 104,4 Punkten ebenfalls sehr gut abschnitten, liegen die Kandidatinnen und Kandidaten, welche als Vorbereitung „nur“ die Test-Broschüre bearbeiteten, mit ihrem Mittelwert von 99,4 Punkten in der Mitte. Wenn C und E gleiche Ergebnisse erzielen, wäre aus dieser Untersuchung kein Vorteil der Trainingskurse gegenüber der Standardvorbereitung abzuleiten. Bei der Bewertung dieser Ergebnisse ist allerdings zu berücksichtigen, dass die Gruppen sicher nicht gleich sind bezüglich der Ausgangssituation, sondern dass Leistungs- wie Motivationsunterschiede die Wahl der Vorbereitung bestimmen.

Deter (1980, 1982) führte ein Experiment durch, bei der eine Gruppe mitberücksichtigt wurde, die sich nur anhand der gratis abgegebenen Test-Broschüre vorbereitet hatte. Gemäss Untersuchungsplan wurden die Testpersonen in fünf Gruppen eingeteilt, die jeweils unterschiedlich auf den Nachtest vorbereitet wurden. Während Gruppe I (n = 96) an dem Vortest - es handelte sich hier um die gleiche Testversion wie beim Nachtest - und anschliessend an einem Training teilnahm, besuchten die Gruppe II (n = 47) nur das Training und die Gruppe III (n = 49) nur den Vortest. Die Gruppe IV (n = 202) ging ohne Vorbereitung an den Nachtest und die Gruppe V (n = 45) studierte die Test-Broschüre als Vorbereitungshilfe. Das Trainingsprogramm umfasste eine Serie von Übungstests, insgesamt 348 Aufgaben, die in 17 Aufgabengruppen zusammengestellt waren. Dieses Material verteilte sich auf zwei Sitzungen von jeweils ca. 3 Stunden. Die Personen, die am Training teilnehmen konnten, erhielten zusätzlich gezielte Instruktionen, spezielle Bearbeitungshinweise, strategische Hilfeleistungen und Erläuterungen der Konstruktionsprinzipien von Aufgaben.

Dadurch sollten sie eine möglichst hohe "test-wiseness" erlangen. Die Resultate (Deter 1980, S. 57) zeigen, dass diejenigen, die am Vortest nicht teilgenommen haben, aber zur Vorbereitung entweder das Training besuchten oder die Test-Broschüre bearbeiteten, gegenüber den Nicht-Vorbereiteten einen Gewinn von 7,9% erzielten - die Trainingskurse erbrachten wiederum keinen nennenswerten zusätzlichen Gewinn gegenüber dem Studium der Test-Broschüre.

Der Einfluss der Vorbereitungsmittel auf die Testergebnisse im Experiment von Deter (1980, 1982).

Gruppe	n	Vortest	Training	Nachtest	Gewinn*
I	96	ja	ja	ja	11,7%
II	47		ja	ja	7,9%
III	49	ja		ja	9,6%
IV	202			ja	0,0%
V	45		Broschüre	ja	7,9%

*Gewinn in % im Vergleich zu den Unvorbereiteten (Gruppe IV)

Ein vom Testkonstrukteur unabhängiges Evaluationsgremium (Bartussek et al. 1984, 1986) untersuchte gleichfalls den Einfluss der Vorbereitungsmöglichkeiten auf die Testleistungen im TMS. Insgesamt 26 Probanden (10 weiblich, 16 männlich) nahmen an einem Vortest, einem Testtraining und am Nachtest teil, während die Kontrollgruppe (n = 33, 16 weiblich, 17 männlich) nur den Vortest und Nachtest bearbeiteten. Die Versuchsgruppe wurde zwischen den beiden Testungen 50 Stunden lang von zwei Diplompsychologen trainiert. Als Resultate hält das Gremium fest (Bartussek et al. 1986, S. 9/11), dass nach einem Studium der Test-Broschüre die zweimalige Bearbeitung unterschiedlicher TMS-Versionen eine durchschnittliche Verbesserung von etwa fünf Punkten erbringt und eine zweimalige Bearbeitung zusammen mit dem Training zu einer mittleren Verbesserung um etwa 15 Punkte führt. Die Hinzufügung des Trainings brachte also eine Verbesserung um zusätzliche zehn Punkte, was ungefähr der Hälfte einer Standardabweichung entspricht. Die statistisch bedeutsame Leistungsvorteile der Versuchsgruppe zeigen sich in den Untertests "Konzentriertes und sorgfältiges Arbeiten" (3,6 Punkte), "Schlauchfiguren" (2,2 Punkte), "Figuren lernen" (2,0 Punkte) und "Textverständnis" (1,6 Punkte). Bei der Kontrollgruppe liess sich nur für den Untertest "Konzentriertes und sorgfältiges Arbeiten" eine statistisch bedeutsame Leistungssteigerung durch den Wiederholungseffekt feststellen. Diese Untersuchung hat vor allem zur

Gestaltung der Test-Broschüre beigetragen. Aufgrund dieser Befunde wurden die dort enthaltenen Trainingshinweise erstellt.

In der Arbeit über die Trainierbarkeit der Leistung im Untertest "Konzentriertes und sorgfältiges Arbeiten" (Fay 1989) zeigte es sich, dass durch blosses Wiederholen der Bearbeitung dieses Subtests eine Leistungssteigerung um bis zu 70% erzielt werden kann. Berücksichtigt die Testperson zusätzlich die in der Test-Broschüre für diesen Untertest abgegebenen Bearbeitungshinweise, so besteht die Möglichkeit, ihre Anfangsleistung zu verdoppeln. Rangplatzverschiebungen sind damit nicht verbunden. Benachteiligt wird somit nur, wer sich nicht vorbereitet hat.

In der Studie von Klieme & Espey (1992) bildeten die Hinweise zum Untertest "Fakten lernen" der Test-Broschüre die Grundlage für das angewandte Training. Die Leistungsgewinne bei 78 Teilnehmern und Teilnehmerinnen lagen im Mittel in der Grössenordnung einer halben Standardabweichung.

Schlussfolgerungen

Bei einer Zusammenfassung der Befunde über den Einfluss von Übung und Training auf die Leistungen in einem standardisierten Test gelangt man zu folgenden Feststellungen (siehe auch Deter 1982, S. 68 ff.):

Testwiederholung führt - im Durchschnitt der Untersuchungsgruppe - zu einer mehr oder weniger ausgeprägten Leistungssteigerung. Der Hinweis der Standard-Vorbereitung, einmal einen Test unter ernstfallnahen Bedingungen durchzuarbeiten, ist deshalb wichtig. In der Schweiz wurde eine veröffentlichte Originalversion in die französische und italienische Sprache übersetzt. Alle drei Sprachgruppen können so auf eine Originalversion als Vorbereitungsgrundlage zurückgreifen.

Der Einfluss der zur Testvorbereitung gratis abgegeben Hilfsmittel auf die Testleistung wurde in einem Experiment untersucht. Dabei erzielten diejenigen, die sich mit diesen Unterlagen vorbereitet haben, im Mittel die gleichen Ergebnisse wie diejenigen, die zusätzlich trainiert wurden.

Einerseits ist es also notwendig, die Vorbereitungsmaterialien genau durchzuarbeiten und die standardisierten Vorbereitungsempfehlungen der Test-Broschüre zu befolgen. Dies bringt nachweislich bessere Leistungen. Zu dieser standardisierten Vorbereitung gehört auch das Durcharbeiten einer vollständigen Version des TMS unter realistischen Zeitbedingungen. Dieser Aufwand bleibt insgesamt in vertretbarem Rahmen.

Andererseits gibt es aus allen Untersuchungen **keine Hinweise, dass ein zusätzliches Training** zur Standardvorbereitung die Leistungen verbessert. Dadurch bleibt dem Test der Vorwurf erspart, dass übermässige zeitliche und finanzielle Ressourcen für ein erfolgreiches Absolvieren notwendig wären.

Sollte eine Testeinführung in der Schweiz erfolgen, kann der Hinweis auf die Vorbereitung entsprechend der Test-Broschüre gegeben werden. Allenfalls wäre es sinnvoll, einen „Probetesttermin" für die Bewerberinnen und Bewerber beispielsweise seitens der Schulen zu organisieren, bei dem ein veröffentlichter Test unter „Ernstfallbedingungen" bearbeitet wird. Es ist erfahrungsgemäss nicht einfach, den Test zu bearbeiten und gleichzeitig auf die Zeitvorgaben zu achten. Dies sollte nur sehr moderate zusätzliche Mittel erfordern.

Weitergehenden Bemühungen, die Kandidatinnen und Kandidaten in Kursen und ähnlichen zusätzlichen Vorbereitungsmassnahmen gegen Entgelt zu trainieren, sollte wo immer nur möglich Einhalt geboten werden.

Literatur

Allalouf, Avi (1996). The effect of coaching on the predective validity of scholastic aptitude tests. Paper presented at the Annual Meeting of the American Educational Research Association, New York, April 1996.

Bartussek, D; Raatz, U; Schneider, B & Stapf, KH (1984). Die Evaluation des "Tests für medizinische Studiengänge". Erster Zwischenbericht. Bonn

Bartussek, D; Raatz, U; Stapf, KH & Schneider, B (1986). Die Evaluation des "Tests für medizinische Studiengänge". Zweiter Zwischenbericht. Bonn

Deidesheimer Kreis (1993). Feststellung der Studieneignung im Rahmen der Hochschulzulassung. Studienfeldbezogene Verfahren zur Feststellung der Studiereignung bei Hochschulzulassungsentscheidungen in Numerus-clausus- und anderen Studienfächern. Bonn: Bericht für das Bundesministerium für Bildung und Wissenschaft.

Deter, Bernhard (1980). Übbarkeit von Leistungen im TMS. In G. Trost et al. Modellversuch "Test für medizinische Studiengänge". Vierter Arbeitsbericht: 1. Januar 1980 bis 31. August 1980. Bonn: Institut für Test- und Begabungsforschung.

Deter, Bernhard (1982). Zum Einfluss von Übung und Training auf die Leistung im "Test für medizinische Studiengänge" (TMS). Braunschweig: Agentur Pedersen.

Fay, Ernst (1985). Vorbereitungsmöglichkeiten auf den Test: Was gibt es? Wie wird es genutzt? Nutzt es? In G. Trost et al. Modellversuch "Tests für medizinische Studiengänge". Auswertungen zum achten und neunten Testtermin und Ergebnisse weiterer Begleituntersuchungen zum Test. Zehnter Arbeitsbericht: 1. April 1984 bis 30. September 1985. Bonn: Institut für Test- und Begabungsforschung.

Fay, Ernst & Freitag, Gerd (1989). Über die Übbarkeit der Leistung im Test "Konzentriertes und sorgfältiges Arbeiten". In G. Trost (Hrsg.) Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 13. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Kirchenkamp, Thomas & Mispelkamp, Harald (1988). Beziehungen zwischen Leistungen im Test für medizinische Studiengänge und verschiedenen Vorbereitungsmassnahmen, Einstellungen zum Vergabeverfahren sowie links- bzw. rechtshändiger Schreibweise. In

G. Trost (Hrsg.) Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 12. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Klieme, Eckhard & Espey, Jürgen (1992). Trainingseffekte beim Merkfähigkeitstest “Fakten lernen”. In G. Trost (Hrsg.) Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 16. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Mispelkamp, Harald (1987). Training für den TMS: Darstellung und Bewertung von Trainingsmaterialien, Bearbeitungshinweisen und formalen Lösungsstrategien. In G. Trost et al. Test für medizinische Studiengänge (TMS): Zehnter und elfter Termin des Übungsverfahrens. Erster Termin im besonderen Auswahlverfahren. Weitere Untersuchungen zum Test. 11. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Nauels, Heinz-Ulrich (1991). Angebote zur Vorbereitung auf den “Test für medizinische Studiengänge”: Neues auf dem Bücher- und Trainingsmarkt. In G. Trost (Hrsg.) Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 15. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Trost, Günter (Hrsg.) (1977-1995). Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 1. - 19. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Zentralstelle für die Vergabe von Studienplätzen (Hrsg.) (1995). Test-Info, Herbst 94. Dortmund: ZVS.

Anforderungen an das Zulassungsverfahren für das Medizinstudium in der Schweiz: Leitlinien für die Entwicklung eines eignungsdiagnostischen Verfahrens

U. Schallberger

1. Einleitung und Vorblick

Geht man davon aus, dass (auch) in der Schweiz demnächst ein Numerus Clausus für die medizinischen Studienfächer unvermeidbar werden könnte, dann stellt sich das Problem eines sinnvollen Zulassungsverfahrens. In den vorangegangenen Beiträgen ist deutlich geworden, dass viele andere Länder zum Teil schon seit langem mit demselben Problem konfrontiert sind und zu dessen Lösung ganz verschiedenartige Verfahren benutzen. Bei der Besprechung dieser Verfahren sind auch eine Reihe von Gütekriterien zur Sprache gekommen, an denen sie zu messen sind. Diese Gütekriterien stammen aus der psychologischen Eignungsdiagnostik. Ziel dieses Beitrags ist es nun, diese eignungsdiagnostische Perspektive als solche etwas herauszuarbeiten und auf die aktuelle Problemlage hier in der Schweiz anzuwenden. Um die zentrale Schlussfolgerung vorwegzunehmen: Es wird sich dabei ergeben, dass in der gegenwärtigen Situation der im Beitrag von K.-D. Haensgen vorgestellte "Eignungstest für das Medizinstudium" (EMS) wohl **die** optimale Lösung darstellt. Um Missverständnissen vorzubeugen: Selbstverständlich gibt es andere mögliche Sichtweisen, die zu einem andern Schluss führen können. Die These ist hier nur, dass – wenn das Zulassungsverfahren überhaupt eignungsdiagnostischen Charakter haben soll – der EMS bei den heutigen Gegebenheiten ein optimaler Kompromiss zwischen Wünschbarem und Machbarem darstellt.

Um die Überlegungen, die zu diesem Schluss führen, nachvollziehbar zu machen, wird wie folgt vorgegangen: Zunächst wird kurz auf den Hintergrund der eignungsdiagnostischen Perspektive eingegangen (Abschnitt 2). Dann werden einige Rahmenbedingungen skizziert, denen das gesuchte Zulassungsverfahren zu genügen hat, wobei sich auch zeigen wird, warum die eignungsdiagnostische Sichtweise in diesem Zusammenhang überhaupt von Bedeutung ist (Abschnitt 3). Die folgenden Abschnitte dienen schliesslich der Erläuterung dieser Sichtweise und ihrer Konsequenzen für das gestellte Problem.

2. Die Allgegenwart von "Eignungsdiagnostik" und der Sinn einer Verwissenschaftlichung entsprechender Verfahren

Versteht man unter Eignungsdiagnostik den Versuch, die Eignung von Menschen für eine Tätigkeit festzustellen (zu "diagnostizieren"), **bevor** die Men-

schen die betreffende Tätigkeit ausüben, dann handelt es sich dabei natürlich um ein ubiquitäres Phänomen. So haben beispielsweise alle Selektionsinstrumente im Bildungssystem (inklusive Notengebung, sofern damit eine Promotions- oder Zulassungsfunktion verbunden ist) eine eignungsdiagnostische Komponente. Genau dasselbe gilt aber auch für die Arbeitswelt. Dort besteht ja seit jeher und heute in extremem Masse ein System im Sinne eines "Numerus Clausus": Sind mehr Bewerber da als Ausbildungs- oder Arbeitsplätze, dann wird mit grösster Selbstverständlichkeit selegiert, unabhängig davon, ob es sich dabei um Lehrstellen handelt oder um höchste Managerpositionen. Dabei werden ganz verschiedene "Instrumente" eingesetzt, vom Studium der Bewerbungsunterlagen über das Vorstellungsgespräch bis hin zu mehrtägigen Testuntersuchungen und den noch aufwendigeren "Assessment Center"-Verfahren. Kernpunkt aller Bemühungen ist dabei, die zu treffende Auswahl unter den Bewerbern hinsichtlich Eignung zu optimieren.

Bei der Legitimation eines konkreten Verfahrens dieser Art lassen sich zwei grundsätzlich verschiedene Vorgehensweisen unterscheiden: Auf der einen Seite stehen Argumentationen nach dem sogenannten "per fiat"-Prinzip, das man alltagssprachlich so umschreiben könnte: "Es ist mir unmittelbar evident, dass das Verfahren die wesentliche Information liefert, also wird es das tun!" Aus der andern Perspektive gilt ein Verfahren erst dann als brauchbar und rational legitimierbar, wenn sich empirisch nachweisen lässt, dass es tatsächlich die Eignung berücksichtigt. Dieser zweite Standpunkt bildet die Grundhaltung der wissenschaftlich fundierten psychologischen Eignungsdiagnostik. Sie kann sich dabei auf eine Fülle von Beispielen berufen, in denen sich "per fiat"-Verfahren bei der empirischen Überprüfung in Langzeitstudien als problematisch, ja als unbrauchbar oder sogar kontraproduktiv erwiesen haben. Die Suche nach den Gründen für solche, für die Beteiligten oft völlig überraschenden Befunde ergab eine Vielzahl von Einsichten in die Schwierigkeiten des eignungsdiagnostischen Unterfangens. Daraus entwickelte sich eine eigene wissenschaftliche Disziplin, die psychologische Eignungsdiagnostik, die auch die theoretischen und methodischen Hilfsmittel für die Entwicklung rational vertretbarer Selektionsverfahren bereitstellt. Das Ganze ist übrigens nicht nur von akademischem Interesse: Die Kosten/Nutzen-Relation eines Selektionsverfahrens lassen sich nur auf diesem wissenschaftlichen Hintergrund wirklich optimieren.[1]

---

[1] Auf den Aspekte des Nutzens eines Selektionsverfahrens wird im folgenden nur am Rande eingegangen. Eine kurze Einführung in die entsprechenden Überlegungen findet sich in Schallberger, U. (1996). Nutzen, Fairness, Validität und Akzeptanz von Selektionsverfahren. In K.-D. Haensgen et al., Eignungsdiagnostik und Medizinstudium (S. 38-42). Bericht 2 des Zentrums für Testentwicklung und Diagnostik an der Universität Fribourg.

Sozusagen als Quintessenz dieser eignungsdiagnostischen Perspektive ergibt sich eine Liste sogenannter Gütekriterien für eignungsdiagnostische Verfahren, die zum Teil im Beitrag von G. Trost vorgestellt wurden. Die meisten dieser Kriterien sind übrigens nicht perfekt erfüllbar. Die Meinung ist aber, dass in jedem Fall objektiv abzuklären ist, in welchem Ausmass sie erfüllt sind, sodass Minus- und Pluspunkte eines Verfahrens offen auf dem Tisch liegen und eine sachliche Diskussion möglich ist. Auf den Begründungszusammenhang und die Relevanz dieser Kriterien wird in Abschnitt 4ff. zurückzukommen sein.

3. Rahmenbedingungen und -anforderungen

Vorerst ist es aber notwendig, sich nochmals die Rahmenbedingungen und -anforderungen zu vergegenwärtigen, denen das in unserem Fall gesuchte Zulassungsverfahren zu genügen hat. Es ist ja wenig sinnvoll, sich Verfahren auszudenken, die im gegebenen Rahmen ohnehin nicht realisierbar sind. Im folgenden werden die fünf wichtigsten Rahmenbedingungen thesenhaft genannt und andiskutiert.

1. **Eine gesamtschweizerische Lösung wäre sinnvoll:** Natürlich gibt es stichhaltige Überlegungen, die für eine Autonomie der Hochschulen auch im Bereich der Zulassung sprechen, beispielsweise der Gedanke, dass eine Eigenverantwortung für die Selektion auch eine grössere institutionelle Verantwortung für den Ausbildungserfolg der (selber ausgewählten) Studierenden impliziert. Zumindest heute, wo sogenannte Umleitungen von Studierenden vor und während des Medizinstudiums zwischen den Universitäten notwendig sind, ist eine dezentrale Lösung aber mit grossen praktischen Schwierigkeiten verbunden.

2. **Eine Gleichbehandlung aller Kandidatinnen und Kandidaten (unabhängig von Herkunftskanton, Schule, Sprache, Geschlecht) ist notwendig:** Dieser zweite Punkt braucht wohl keine weitere Begründung. In unserem Bildungssystem ist es zur Zeit zum Beispiel kaum denkbar, dass – wie in den USA – die (irgendwie bestimmte) Selektivität der besuchten Mittelschule zum Zulassungskriterium gemacht würde. Die Idee der Gleichbehandlung spricht übrigens auch gegen die Verwendung von Maturanoten oder Wissenstests. Die Notengebungspraxis variiert ja zwischen den Schulen und Landesteilen sehr – und würde vermutlich noch mehr variieren, wenn die Maturanoten eine selektive Bedeutung bekämen. Anderseits würde sich im Gefolge der Einführung von Wissenstests, die sich von der Natur der Sache her auf etwas beziehen, das bei entsprechendem Einsatz erwerbbar ist, relativ rasch eine "Trainingsindustrie" herausbilden, von der nur jene Kandidaten profitieren können, die über die notwendigen finanziellen Mittel verfügen.

3. **Die zeitlichen und finanziellen Ressoucen für die Entwicklung eines Zulassungsverfahrens sind äusserst beschränkt:** Betrachtet man ausländische Beispiele von Hochschulzulassungsverfahren oder Beispiele von Selektionsverfahren in anderen Anwendungskontexten wird deutlich, dass die problembewusste Entwicklung eines derartigen Verfahrens ausserordentlich aufwendig ist. Beispielsweise sei daran erinnert, dass die aktuelle Version des amerikanischen MCAT auf 8 Jahren intensiver Vorarbeiten und Forschung beruht (siehe Beitrag von J. L. Hacket). Ähnliches liesse sich über andere Verfahren sagen. Offensichtlich hat die Schweiz gegenwärtig weder die Zeit noch die Mittel, um eine völlig eigenständige Lösung von Grund auf zu erarbeiten.

4. **Die verfügbaren Ressourcen für die Durchführung des Zulassungsverfahrens sind beschränkt:** Auch dieser Punkt begrenzt unseren Handlungsspielraum. Angesichts der knappen Ressourcen ist es zum Beispiel nicht realistisch, individuelle Selektionsgespräche mit allen Studieninteressenten durchführen zu wollen. Realisierbar ist höchstens ein Verfahren, in dem viele Bewerber gleichzeitig getestet bzw. untersucht werden können, wobei individuelle Interviews vielleicht für eine beschränkte Zahl von Sonder- bzw. Grenzfällen denkbar sind.

5. **Das neue Zulassungssystem soll den Gesichtspunkt der Eignung besser berücksichtigen als dies im heutigen System der Selbstselektion der Fall ist:** Der heutige Zugang zum Medizinstudium basiert – eine bestandene Matur vorausgesetzt – auf **Selbstselektion**: Wer am Medizinstudium interessiert ist und sich selber dazu als fähig erachtet, wird zugelassen. Die sogenannte **Fremdselektion** setzt erst in den propädeutischen Prüfungen ein. Soweit zu erkennen ist, gehen alle interessierten Kreise davon aus, dass das neue Zulassungsverfahren über die Selbstselektion hinaus noch eine objektivierte Berücksichtigung der Eignung der Studienbewerber beinhalten soll. Würde diese Forderung nicht erhoben, gäbe es an sich eine perfekte Lösung des Zulassungsproblems, nämlich das **Lotterieverfahren**, die Auswahl per Zufall. Dieses Vorgehen würde nämlich den vier oben genannten Anforderungen geradezu maximal gerecht (gesamtschweizerisch durchführbar, garantierte Gleichbehandlung, äusserst geringe Verfahrensentwicklungs- und -durchführungskosten). Wie die Übersicht von G. Trost gezeigt hat, spielt aber dieses Verfahren international kaum eine Rolle – verständlicherweise: Bereits der gesunde Menschenverstand sträubt sich ja gegen die Delegation des Zulassungsentscheids an das Los. Die Überlegung ist die: Wenn die Studienplätze schon beschränkt sind, dann erscheint es – aus der Sicht der Öffentlichkeit, der Hochschulen und der Bewerber – gerecht und vernünftig, jene Kandidatinnen und Kandidaten systematisch zu bevorzugen, die über eine bessere Eignung verfügen (wobei natürlich unter Eignung verschiedenes verstanden werden kann). Das Prinzip der Gleichbehandlung wird dadurch

nicht aufgegeben, sondern dahingehend modifiziert, dass Bewerber mit gleicher Eignung Anspruch auf Gleichbehandlung haben sollen. Meist wird dieses modifizierte Gleichbehandlungsprinzip als "Fairness" bezeichnet.

Wenn nun aber die Eignung als Zulassungskriterium berücksichtigt werden soll, dann wird das Zulassungsverfahren automatisch zu einem eignungsdiagnostischen Verfahren im früher umschriebenen Sinne - mit allen Schwierigkeiten, die mit einem solchen Verfahren verbunden sind.

4. Zur Grundstruktur eignungsdiagnostischer Aussagen

Um die Quellen der Schwierigkeiten der Eignungsdiagnostik und die Möglichkeiten für deren Bewältigung in den Blick zu bekommen, ist zunächst auf die Grundstruktur einer eignungsdiagnostischen Aussage einzugehen.

Eine eignungsdiagnostische Aussage hat von ihrer Natur her den Charakter einer **Prognose**: Aufgrund gegenwärtig verfügbarer Informationen, **Prädiktoren** genannt, wird eine Aussage über das Ausmass des künftigen Erfolgs, **Kriterium** genannt, gemacht. Eine solche Aussage ist in jedem Fall – genauso wie die Prognosen der Meteorologen oder der Konjunkturforscher – mit Unsicherheiten behaftet, das heisst eine Wahrscheinlichkeitsaussage. Die Quellen der Unsicherheit sind dabei höchst unterschiedlicher Natur. Abbildung 1 soll dies illustrieren.

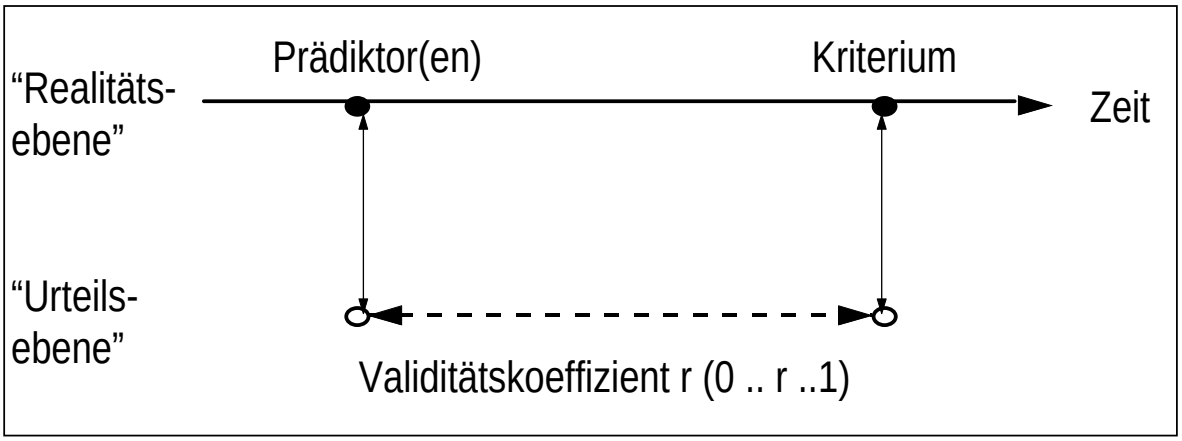


Abb. 1: Zur Struktur einer eignungsdiagnostischen Aussage

In dieser Abbildung repräsentiert die Horizontale die Zeitachse, auf der Prädiktor(en) und Kriterium lokalisiert sind. In der Vertikalen sind zwei Ebenen unterschieden, die "Realitätsebene" und "Urteilsebene" genannt werden. Das Gemeinte lässt sich am einfachsten anhand eines Beispiels illustrieren:

Angenommen, jemand gehe davon aus, das zu prognostizierende Kriterium sei "guter Arzt", und als Prädiktor solle "Einfühlungsvermögen" verwendet werden, und zwar aufgrund der Auffassung, dass ein guter Arzt über ein Minimum an Einfühlungsvermögen für die Sicht und die Sorgen des Patienten verfügen müsse. Diese Überlegungen betreffen im Schema die "Realitätsebene": Sie beinhalten Aussagen bzw. besser: Annahmen über die Realität. Um sie für ein eignungsdiagnostisches Verfahren fruchtbar machen zu können, müssen diese Aussagen aber noch in die im Schema so genannte "Urteilsebene" transformiert werden. Das heisst, es muss ein Verfahren gewählt werden, das erlaubt, ein Urteil über das Einfühlungsvermögen konkreter Kandidaten zu fällen; erst ein solches Urteil kann ja tatsächlich als Prädiktor verwendet werden. Wenn auch noch die Güte der Prognose überprüft werden soll (und das **muss** sie aus eignungsdiagnostischer Perspektive immer, siehe oben, Abschnitt 2), dann ist auch noch ein Beurteilungsverfahren zu entwickeln, das erlaubt, zwischen guten und weniger guten Ärzten ("Kriterium") zu unterscheiden.

Sind Prädiktor(en) und Kriterium in diesem Sinne in die "Urteilsebene" übersetzt, erlaubt dies nach einer gewissen Zeit, den Prädiktor mit dem Kriterium empirisch in Beziehung zu setzen. Das entsprechende Ergebnis misst die Güte der Vorhersage, den Grad der Unsicherheit bzw. das Ausmass der Fehlurteile, die mit dem gewählten Vorgehen verbunden sind. Der entsprechende Koeffizient wird **prognostische Validität** genannt und ist in mehreren der vorangegangenen Beiträge bereits benutzt worden. Dieser Koeffizient bewegt sich in der Regel zwischen 0 und 1. "1" bedeutet eine perfekte Voraussage – ein Ziel, das real nie erreicht werden kann. "0" bedeutet, dass die "Güte" der Prognose genau derjenigen des Lotterieverfahrens entspricht. Ein negativer Koeffizient würde übrigens dann resultieren, wenn das Zulassungsverfahren systematisch die weniger Geeigneten bevorzugt, was zwar gänzlich unerwünscht ist, in der Realität aber doch immer wieder vorkommt.

5. Determinanten der prognostischen Validität einer eignungsdiagnostischen Aussage

Das Schema in Abbildung 1 erlaubt nun auch, die Determinanten der prognostischen Validität einer eignungsdiagnostischen Prognose zusammenzustellen. Diese Determinanten sind unabhängig vom genannten Beispiel; sie gelten für beliebige Prädiktoren und Kriterien, die zur Lösung irgendeines eignungsdiagnostischen Problems herangezogen werden.

1. **Faktische Rolle des Prädiktors für das Kriterium auf der Realitätsebene:** Es stellen sich hier Fragen wie: Gilt der postulierte Zusammenhang zwischen Prädiktor und Kriterium tatsächlich? Spielen nicht andere Faktoren auch noch eine, eventuell sogar eine grössere Rolle? Es dürfte unmittelbar einsichtig

sein, dass die Anzahl und die Effektstärke solcher anderer Faktoren die prognostische Validität eines gewählten Prädiktors negativ beeinflussen.

2. **Zeitabstand zwischen Prädiktor und Kriterium:** Weil sich der Mensch in der einen wie in der andern Richtung entwickeln kann, ist naturgemäss die prognostische Validität generell umso geringer, je grösser der Zeitabstand zwischen Prädiktor und Kriterium ist. Im Falle der Selektion für das Medizinstudium ist dieser Punkt besonders problemhaltig, weil es sich um junge Menschen handelt, die sich zudem in einer Ausbildung befinden, die – wenn sie entsprechend angelegt ist – selber noch zu einer Reihe von Entwicklungen Anlass geben könnte.

3. **Objektivität, Zuverlässigkeit und Validität der Erfassung des Prädiktors:** Der zentrale Punkt ist hier die Validität des Urteils über den Prädiktor, die von der prognostischen Validität zu unterscheiden ist. Es besteht aber ein Zusammenhang: Je geringer diese "Urteils-Validität" ist, umso geringer ist die prognostische Validität des Verfahrens. Im obigen Beispiel: Wenn das Einfühlungsvermögen eines Menschen nicht valide erfasst wird, dann kann das entsprechende Urteil selbstverständlich auch keine substantielle prognostische Validität haben. Die beiden andern Kriterien, Objektivität und Zuverlässigkeit (sie wurden im Beitrag von G. Trost definiert), gewinnen ihre Bedeutung dadurch, dass sie zwar nicht hinreichende, aber notwendige Bedingungen der Urteilsvalidität benennen: Beispielsweise kann eine Beurteilung, bei der sich zeigen lässt, dass sie stärker vom Beurteiler als vom Beurteilten abhängt (= fehlende Objektivität), a priori nicht valide sein.
   Diese Überlegungen erlauben übrigens, verschiedene Verfahrenstechniken gegeneinander abzuwägen, wie es im Beitrag von G. Trost ansatzweise geschah. So ist zum Beispiel aus vielen Kontrolluntersuchungen bekannt, dass Beurteilungen aufgrund von frei geführten Gesprächen (Interviews) höchstens eine sehr geringe Objektivität und Zuverlässigkeit, und damit auch eine geringe prognostische Validität haben. Erst relativ strenge Standardisierungsmassnahmen (vorgegebene Fragen, vorgegebene Beurteilungsschemata, Interviewerschulung) erlauben, aus Interviews potentiell brauchbare Prädiktoren abzuleiten.

   Analog können auch verschiedene Kategorien von Prädiktoren evaluiert werden. Beispielsweise hat es sich als sehr viel schwieriger erwiesen, objektive, zuverlässige und valide Beurteilungen von Charaktereigenschaften zu gewinnen als dies für Fähigkeiten der Fall ist. Am günstigsten sind die Aussichten, wenn es sich um relativ verhaltensnahe Merkmale handelt.

4. **Objektivität, Zuverlässigkeit und Validität der Erfassung des Kriteriums:** Dasselbe, was eben für die Prädiktoren gesagt wurde, gilt mutatis mutandis auch für das Kriterium. Es macht aus eignungsdiagnostischer Sicht

wenig Sinn, ein Kriterium voraussagen zu wollen, das nicht in einigermassen objektiver, zuverlässiger und valider Weise erfasst werden kann, weil damit die entsprechenden Prognosen praktisch unüberprüfbar werden. Eine Legitimation des Verfahrens ist unter solchen Umständen eine reine Glaubensfrage.

Vergegenwärtigt man sich nun alle diese Determinanten der prognostischen Validität eines Selektionsverfahrens, lässt sich vielleicht erahnen, warum die Entwicklung eines solchen Instruments derart aufwendig ist, wie es weiter oben erwähnt wurde. Jeder der genannten Aspekte bedarf sorgfältiger empirischer Abklärungen, bevor ein Verfahren in einem rationalen Sinne legitimierbar ist.

Abschliessend zu diesen Überlegungen sei noch eine Vorgehensweise erwähnt, mit deren Hilfe manche Schwierigkeiten im Zusammenhang mit den Punkten 1, 3 und 4 zwar in keiner Weise eliminiert, aber wenigstens etwas abgedämpft werden können. Es geht dabei um den Unterschied zwischen sogenannten **eigenschaftsorientierten** und **simulationsorientierten** Verfahren. Worin dieser Unterschied besteht, sei am Beispiel der Pilotenselektion illustriert: Eigenschaftsorientiert wäre ein Verfahren, das als Prädiktoren Eigenschaften benutzt, von denen angenommen wird, dass sie für einen "guten Piloten" notwendig sind (zum Beispiel räumliches Vorstellungsvermögen, Konzentrationsfähigkeit, Stressresistenz etc.). Die Beurteilung der Eigenschaftsausprägungen erfolgt mittels pilotenunspezifischen Standardinstrumenten. Simulationsorientiert wäre hingegen ein Verfahren, das möglichst direkt von den Anforderungen ausgeht, wie sie im Pilotenalltag tatsächlich auftreten, zum Beispiel hinsichtlich Konzentrationsleistungen oder Stressresistenz. Als Prädiktoren werden dann Leistungen in Testverfahren benutzt, die diese spezifischen Anforderungen möglichst gut simulieren. "Möglichst gut" bezieht sich dabei nicht auf den Augenschein, sondern auf die psychologische Strukturverwandtschaft von simulierter und Realsituation. Nach allen Regeln der Kunst konstruierte simulationsorientierte Verfahren sind im allgemeinen prognostisch valider als eigenschaftsorientierte Verfahren, und zwar primär aus zwei Gründen: Zum einen ist der Zusammenhang zwischen Prädiktor und Kriterium a priori weniger spekulativ. Zum andern ermöglicht dieses Vorgehen den interessierenden Prädiktor und das Kriterium in einer verhaltensnäheren Form zu formulieren. In der Sprache von Abbildung 1 ausgedrückt: Die Distanz von "Urteilsebene" und "Realitätsebene" und die damit verbundenen Probleme werden minimiert.

6. Schlussfolgerungen

Ausgangspunkt unserer Überlegungen war das Problem, dass auch in der Schweiz demnächst ein Numerus Clausus in den medizinischen Fächern notwendig werden könnte und dass sich deswegen die Frage nach einem sinnvollen Zulassungsverfahren stellt. Da von diesem Verfahren allgemein erwartet wird,

dass es auch eine eignungsdiagnostische Funktion haben soll, wurden anschliessend einige zentrale Probleme besprochen, die mit einer solchen Zielsetzung verbunden sind. Konfrontiert man nun das Gesagte mit der aktuellen schweizerischen Situation, zeigt sich rasch, dass aus eignungsdiagnostischer Sicht im Moment keine brauchbare Alternative zu dem im vorigen Beitrag vorgestellten "Eignungstest für das Medizinstudium" (EMS) in Sicht ist. Zusammenfassend führen vor allem folgende Punkte zu diesem Schluss:

- Die problembewusste Entwicklung eines eigenen eignungsdiagnostischen Verfahrens würde zeitliche und materielle Ressoucen voraussetzen, die gegenwärtig in der Schweiz nicht zur Verfügung stehen. Es ist daher notwendig, an ausländischen Vorarbeiten anzuknüpfen. Um die Chance zu optimieren, dass es auch in der Schweiz im gewünschten Sinne funktioniert, empfiehlt es sich, ein Instrument zu wählen, das aus einem Land mit einem möglichst ähnlichen Hochschulsystem stammt. Der EMS – in Deutschland im Verlaufe vieler Jahre entwickelt und überprüft – ist das einzige vorhandene Verfahren, das diesem Kriterium genügt.

- Der EMS erfüllt die skizzierten Rahmenanforderungen optimal: Er ist rasch verfügbar, gesamtschweizerisch anwendbar und ökonomisch durchführbar. Es gibt ferner empirische Belege dafür, dass er weitestgehende Fairness garantiert.

- Der EMS verwendet ein – für ein eignungsdiagnostisches Verfahren – realistisches Erfolgskriterium, nämlich den Studienerfolg. Damit liegt auch der Zeitabstand zwischen Prädiktion und Kriterium in einem vernünftigen Rahmen. Natürlich wäre aus übergeordneter Perspektive ein Kriterium vorzuziehen, das die spätere Berufstätigkeit des Arztes betrifft. Das prognostische Problem würde dadurch aber viel komplexer. Es wäre – fast unabhängig vom Aufwand – mit einer prognostischen Validität zu rechnen, die jene eines Lotterieverfahrens kaum substantiell übertreffen würde.

- Der EMS ist ein aufgrund umfangreicher Vorstudien konstruiertes simulationsorientiertes Verfahren. Er zielt darauf ab, als wichtig identifizierte kognitive Anforderungen, mit denen sich die Studierenden im Verlauf des Studiums konfrontiert sehen, zu simulieren. Der mit vielen Fallstricken versehene "Umweg" über Annahmen, welche notwendige Eigenschaften der Bewerber betreffen, wird dadurch weitgehend überflüssig.

- Entsprechend ist die prognostische Validität des EMS an der oberen Grenze dessen, was bei einer derartigen Selektionsaufgabe und bei vergleichbarem Selektionsaufwand überhaupt zu erreichen ist. Die deutschen Validitätskoeffizienten liegen bei ca. .5, ein Wert, der nach international üblichen Standards

als hoch zu bezeichnen ist und bei vergleichbaren Problemlagen und gleicher Berechnungsmethode kaum je wesentlich übertroffen wird.

Ein nicht zu übersehender Schwachpunkt des EMS besteht in der Tatsache, dass Schweizer Kontrolluntersuchungen, insbesondere zur prognostischen Validität, noch weitgehend fehlen. Der Verweis auf die deutschen Ergebnisse darf ja nicht zum "per fiat"-Argument verkommen. Sollte der EMS tatsächlich eingesetzt werden, ist es daher unumgänglich, dass dieser Einsatz von sorgfältig konzipierten empirischen Untersuchungen begleitet wird – im analogen Falle in der Arbeitswelt heute bei problembewussten Firmen eine Selbstverständlichkeit. Dies setzt voraus, dass die dafür notwendigen Mittel zur Verfügung gestellt werden. Nur mit Hilfe derartiger Begleituntersuchungen wird es möglich sein, objektiv zu bestimmen, ob und in welchem Ausmass das gewählte Verfahren einem Entscheid durch das Los überlegen ist. Aus eignungsdiagnostischer Sicht ist es anderseits nur der Grad dieser Überlegenheit, der den ganzen Aufwand rechtfertigt und der Hochschule wie auch den vom Numerus Clausus betroffenen Studienanwärtern einen Nutzen bringt.

Adressen der Autoren

Dr. John L. HACKETT

MCAT Section
Association of American Medical Colleges; 2450 N Street, NW;
Washington, DC 20037
U. S. A.

PD Dr. Klaus-Dieter HÄNSGEN; lic.phil. Rainer HOFER

Zentrum für Testentwicklung und Diagnostik
am Psychologischen Institut der Universität Freiburg
Rte d'Englisberg 9
CH-1763 Granges-Paccot

Prof. Dr. Widar HENRIKSSON; Prof. Dr. Ingemar WEDMAN

Department of Educational Measurement
University of Umeå
S-Umeå 90187

Prof. Dr. Piet J. JANSSEN

Centrum voor Schoolpsychologie
Departement Psychologie
Katholieke Universiteit Leuven
Tiensestraat 102
B-3000 Leuven

Prof. Dr. Urs SCHALLBERGER

Psychologisches Institut
Universität Zürich
Schönberggasse 2
CH-8001 Zürich

Dr. Günter TROST

Institut für Test- und Begabungsforschung
Koblenzerstrasse 77
D-53177 Bonn