UNIVERSITÉ DE FRIBOURG SUISSE
UNIVERSITÄT FREIBURG SCHWEIZ

ZENTRUM FÜR TESTENTWICKLUNG UND DIAGNOSTIK AM DEPARTEMENT FÜR PSYCHOLOGIE
CENTRE POUR LE DEVELOPPEMENT DE TESTS ET LE DIAGNOSTIC AU DEPARTEMENT DE PSYCHOLOGIE
RTE D'ENGLISBERG 9 1763 GRANGES-PACCOT ☎ +41(26)3007986 🖷 +41(26)3009763
WWW.UNIFR.CH/ZTD

Stellungnahme zum Bericht „Evaluation der Eignungstests für das Medizinstudium in Österreich" von Chr. Spiel u.a. im Auftrag des bm:wf vom Januar 2008

Klaus-D. Hänsgen 29.2.2008

Leistungsunterschiede zwischen Frauen und Männern in beiden Eignungstests für das Medizinstudium (Graz bzw. Wien/Innsbruck) sowie der Unterschied von Österreicherinnen und Österreichern gegenüber den Angehörigen der EU-Quote waren nach unserer Kenntnis vor allem Anlass für diese Begutachtung. Da letztgenannter Unterschied durch die Quotenregelung bei der Zulassung nicht wirksam wird, bleibt der „**Gendereffekt**" als bei der Zulassung nicht ausgeglichen bestehen. Auch die beteiligten Universitäten nahmen und nehmen diese Fragen sehr ernst (siehe Bericht zum EMS-AT 2007[1] und 2008 in Vorbereitung), um **testbedingte Benachteiligungen** in jedem Falle zu vermeiden. Bisher konnte eine testbedingte Benachteiligung, die auszugleichen wäre, nicht nachgewiesen werden.

Die Diskussion zum Gutachten soll **so konstruktiv wie möglich** aufgegriffen und um bisher nicht berücksichtigte Aspekte ergänzt werden – **zumal vielen der vorgeschlagenen konkreten Massnahmen vorbehaltlos zugestimmt werden kann.**

Die Begutachtung geht über Gender-Fragen hinaus und versucht eine generelle Beurteilung der Qualität des EMS. Besonders in diesem Teil müssen Besonderheiten der Konstruktion des EMS nachhaltiger berücksichtigt werden. Es sollte in der weiteren Diskussion auch **unterschieden** werden, welche Faktoren für den **Gendereffekt** verantwortlich sind oder gemacht werden und **welche Kritiken geäussert worden sind, die – selbst wenn sie zuträfen – diesen Effekt nicht wirklich erklären. Hier sollte jede Spekulation vermieden werden.**

Prognosekraft als „ultimatives" Kriterium (S. 11)

Wir stimmen voll zu, dass eine hohe **Prognosekraft bezüglich des Studienerfolgs das eigentliche Kriterium** der Brauchbarkeit für den Eignungstest darstellt. Würde dieses Ziel nicht erfüllt, wäre jeder Testeinsatz verfehlt und man könnte stattdessen losen. Der Test ist ein wettbewerbsorientiertes Reihungsverfahren, es sollen diejenigen bevorzugt einen Studienplatz erhalten, die befähigt sind, das Studium in der vorgesehenen Zeit abschliessen und die begrenzten Kapazitäten wieder für die Nächsten frei machen sowie die besseren Studienergebnisse erzielen. Dass die Kapazitäten beschränkt sind und nicht mehr alle Interessenten zugelassen werden können, ist **Ursache – nicht Folge** des Testeinsatzes. Da wesentlich mehr Interessenten als Studienplätze vorhanden sind (4-5 Bewerbungen auf einen Platz), **könnten auch Abge-**

[1] http://www.eignungstest-medizin.at/cms/index.php?option=com_content&view=article&id=10&Itemid=8

lehnte das Studium bewältigen – aber entweder in längerer Zeit oder mit schlechteren Leistungen, wenn der Test richtig funktioniert.

Da der Test 2006 erstmalig in Österreich angewendet wurde, können zeitlich bedingt bisher nur die Ergebnisse der SIP-1 zur Analyse der Studienleistungen verwendet werden. Schuler und Mitarbeiter[2] haben eine sehr umfassende internationale Metaanalyse aller Zulassungskriterien vorgelegt, die auch als **Benchmarking** dienen kann, welche Prognosekraft ein gutes Zulassungsverfahren für Studienerfolg überhaupt erreichen kann und muss. Schulnotendurchschnitte und fachspezifische Studierfähigkeitstests teilen sich mit Korrelationen zwischen 0.40 und 0.50 den ersten Platz – die Fairness von Schulnoten für den **Einzelfall** wird z.B. dann relativiert, wenn Gruppen unterschiedlich streng benotet werden (wofür es bekanntlich Anzeichen gibt). Eignungstests haben demgegenüber den Vorteil, dass die Bedingungen für alle Personen vergleichbar und frei von subjektiven Bewertungseinflüssen gestaltet werden können.

Die Korrelationen des SIP-1-Punktwertes (Wien) mit der EMS-Testleistung sind wie folgt („korrigiert" bezieht sich auf das übliche Verfahren zum Ausgleich der Varianzverringerung beim Testwert durch die Zulassung der Besten):

Gesamt	**0.42**	**(p < .000)**	**korrigiert 0.53**
Männer	0.40	(p < .000)	korrigiert 0.50
Frauen.	0.41	(p < .000)	korrigiert 0.53

Von Interesse ist auch, dass die Mittelwerte der **Nicht Erfolgreichen** in der SIP-1 (unter 67 Punkten) sich **im EMS** für Frauen (106.3) und Männer (106.7) **nicht signifikant** unterscheiden. Die „Erfolgsgrenzen" sind somit vergleichbar, die „Hürde" ist gleich. Es gibt keinen Hinweis auf systematische Unterschätzung des prognostizierten Studienerfolges für Frauen durch den EMS. Eine genauere Darstellung erfolgt im Bericht 2007 des EMS-AT (in Vorbereitung).

Weitergehende Validierungen stehen bisher nur aus Deutschland und der Schweiz zur Verfügung. Die jüngste Validierung in der Schweiz[3] ist dabei besonders hervorzuheben- **weil hier bereits umfassende Studienreformen stattfanden**. Auch für die 2. Vorprüfung betragen die Prognose-Korrelationen 0.45. Der Prozentsatz der Personen, welche die 2. Vorprüfung bestehen, hat sich in Universitäten mit Zulassungsbegrenzung erhöht – nähert sich z.B. in Bern 90%.

Genderunterschiede in den SIP-1 Prüfungen werden korrekt vorhergesagt (S.1)

Von allen Seiten unbestritten ist die Tatsache, dass Frauen in den SIP-1-Prüfungen unter **vergleichbaren** Bedingungen (erste Antritte) schlechtere Leistungen erreichen als Männer (im Gutachten wird Mitterauer u.a. 2007 zitiert, auch früheren Untersuchungen von Frischenschlager u.a. 2005 kamen zu dem Ergebnis). Für Frauen wird festgestellt, dass mit einem Jahr Zeitverlust die Unterschiede aufgeholt werden – es bleibt aber festzuhalten, dass zu Studienbeginn besagte Unterschiede objektiv vorhanden sind. Der EMS kann seinerseits nur den Studienerfolg unter **vergleichbaren** Bedingungen vorhersagen. Insofern entspricht der im Test festgestellte Genderunterschied genau dem in der SIP-1 festgestellten Unterschied und die Studienerfolgsprognose ist insgesamt richtig. Der Test würde sogar falsch prognostizieren, wenn sich dieser Unterschied bei den Prüfungen nicht im Testergebnis widerspiegelt. Dass die Prognosekorrelation für die Geschlechter gleich ist (s.o.), spricht für gleichartige Zusammenhänge zwischen Testergebnis und Studienerfolg.

Sollte es bei Frauen eine Gruppe geben (es sind ja nicht alle), die eine längere Anlaufzeit benötigen – etwa auch, um bestimmte Defizite bei der Aneignung naturwissenschaftlicher Kenntnisse auszugleichen – wird man trotzdem akzeptieren müssen, dass diese in einem objektiven Test

---

[2] Hell, Trappmann, Weigand, Hirn und Schuler (2005): Die Validität von Prädiktoren des Studienerfolges. Eine Metaanalyse. Präsentation als Vorabdruck zum Buch: Studierendenauswahl und Studienentscheidung. Göttingen: Hogrefe (2008)

[3] http://www.unifr.ch/ztd/ems/emseval07.pdf

vor Studienbeginn, dem besagten „wettbewerbsorientierten Reihungsverfahren“ schlechter abschneiden und ein Ausgleich „weil das ja später anders wird“ **wäre nur auf politischem Wege möglich**. Dies wäre aber auch nicht problemlos, weil dann **Fairness im Einzelfall** nicht mehr gegeben wäre: Wenn die Kapazität gleich bleibt, müssten man zum Ausgleich Männer **nicht** zulassen, die die Prüfung laut Prognose in der Realität auch eher bewältigen würden.

Was ist Zuverlässigkeit, was muss zuverlässig sein? (S.8)

Richtig ist, dass die verwendeten (!) Werte eines Tests zuverlässig (reliabel) sein müssen. Im Gutachten wird ein Wert von 0.87 gefordert – die für Eignungsdiagnostik massgebliche DIN 33430 bzw. gleichlautende ÖNORM D 4000 fordern je nach Verfahrensklasse 0.70 bis 0.85 (weil man das so generell nicht festlegen kann).

Beim EMS beträgt die Reliabilität des für die Zulassung verwendeten Testwertes 2007 **0.90 (Schweiz) oder 0.93 (Österreich),** berechnet mittels Testhalbierungsmethode auf Itemebene. Vergleichbare Werte wurden in allen Vorjahren erreicht. Der verwendete EMS-Testwert ist ohne Wenn und Aber ausreichend zuverlässig.

Die Reliabilitäten der einzelnen Aufgabengruppen würden nur Bedeutung erlangen, wenn diese Werte einzeln diagnostisch verwendet werden – etwa im Rahmen einer Bildungsberatung, um Stärken und Schwächen in Bereichen festzustellen. Dafür ist der EMS weder gedacht noch geeignet. Die etwas geringeren Reliabilitäten der Aufgabengruppen sind der „Preis“ für eine grössere Vielfalt, grössere Heterogenität innerhalb der einzelnen Untertests. Diese war den Konstrukteuren des TMS wichtig (mehr siehe unter: *„Verrechnungsfairness“ und Bemerkungen zur Konstruktion*), um durch Vielfalt z.B. die Trainierbarkeit zu minimieren. Indem die Reliabilität des Testwertes „trotzdem“ sehr hoch ist und vor allem die Prognosegüte erwartet hoch ausfällt, wird dieses Vorgehen ebenfalls ohne Einschränkungen legitimiert[4].

Wo sind die Geschlechterunterschiede grösser - Graz oder Innsbruck/Wien? (Seite 3)

Auf Seite 3 unten wird festgestellt, dass die Geschlechterunterschiede für Wien/Innsbruck noch deutlicher seien als für Graz. Diese Feststellung ist für die Anzahl der Zulassungen korrekt (Abb. 1 und 2). Rechnet man allerdings die Genderunterschiede der Tab. 2 und 3 in Prozent der Standardabweichung um, ergibt sich für Wien/Innsbruck ein Unterschied von 39%, für Graz von 43% derselben - was eine gegenteilige Aussage nahelegen würde.

Sind Multiple-Choice-Aufgaben mit nur einer richtigen Lösung wirklich schlechter? (S.9)

Andere Antwortformate (mehrere richtige Lösungen pro Aufgabe) sind denkbar. Für die erprobten vorhandenen Aufgaben ist allerdings keine nachträgliche Änderung möglich, ohne eine erneute empirische Prüfung unter Ernstfallbedingungen notwendig zu machen, weil sich die Schwierigkeit drastisch ändert.

Es ist ein Charakteristikum des EMS, nur ausreichend vorerprobte Aufgaben zu verwenden (das könnte auch die festgestellten Unterschiede zum ähnlichen Untertest „Textverständnis“ in Graz erklären). Ein anderes Vorgehen wäre allerdings nicht problemlos: Es ist für die Rechtsfähigkeit des EMS bedeutsam, dass es pro Aufgabe tatsächlich eine richtige Lösung gibt, die sich ausreichend von allen anderen unterscheidet. Es dürfen z.B. keine sogenannten „Doppeldeutigkeiten“ auftreten, dass falsche Lösungen dennoch als vermeintlich richtig abgeleitet werden könnten. Bisher konnte in diesen Fällen immer der eindeutige empirische Beweis erbracht werden, dass die Leistungsbesten in einem Aufgabenbereich die richtige Lösung auch gehäufter wählen - die Trennschärfe der richtigen Lösung sich von der aller Falschlösungen deutlich genug unterschei-

[4] In der Psychologie gilt: ein nicht reliabler (zuverlässiger) Test kann nicht valide (gültig – hier für die Erfolgsprognose) sein. Andererseits ist muss ein ausreichend valider Test automatisch auch reliabel sein. Zu beachten ist auch, dass es keine Reliabilität „an sich“ gibt, sondern verschiedene Aspekte und Schätzmethoden existieren.

det. Da sich bei mehr richtigen Lösungen pro Aufgabe deren Einzelschwierigkeiten auch unterscheiden, würde dieser Unterschied graduell „aufgeweicht“ und die o.g. Beweisführung zumindest erschwert, wenn nicht unmöglich. Das wäre bei neuen Überlegungen (die möglich sind) zumindest zu bedenken.

Die Literatur ist nach unserer Kenntnis zwar voll von Vermutungen, aber bis heute **eindeutige** empirische Belege dafür schuldig geblieben, dass unterschiedlich viele Lösungsoptionen pro Aufgabe deren **Zuverlässigkeit** tatsächlich verbessert. Alle grossen Studierfähigkeitstests arbeiten, was die Zulässigkeit nur einer richtigen oder besten Lösung betrifft, offenbar nicht ohne Grund nach dem gleichen Prinzip wie der EMS.

Der Einfluss der **Ratewahrscheinlichkeit** wird im Übrigen durch die mehrfach gegebene Instruktion ausgeglichen, alle Fragen zu beantworten – also am Ende immer bei jenen zu raten, die unbeantwortet sind. Damit sind die Bedingungen für alle Personen gleich und echte Leistung wird als Differenz zur Ratewahrscheinlichkeit aufgefasst.

Kann man den Test verkürzen? (S. 11)

Auch die Länge des EMS wird im Gutachten kritisiert. Eine Verkürzung ist grundsätzlich möglich, sogar ohne drastische Einbussen der Prognosekraft. Sowohl in Deutschland als auch in der Schweiz hat man sich aber bisher bewusst für die aktuelle Länge entschieden:

Oft wird kritisiert, dass der EMS die Berufseignung nicht berücksichtige. Diese ist nun nicht nur durch Empathie, sondern auch durch Belastbarkeit, Stressresistenz, Ausdauer und stabile Leistungsfähigkeit über eine längere Zeit (man denke etwa an eine Operation) gekennzeichnet. Auch im Studium werden Anforderungen gestellt, die solche Eigenschaften fordern. Indem der EMS ein volles Eintages-Assessment darstellt, spielen diese Merkmale für eine erfolgreiche Absolvierung auch eine Rolle. Ist der Test zu kurz, verlieren diese Faktoren an Bedeutung.

Es ist des Weiteren in den Evaluationen von 2001 auch nachgewiesen, dass die Einbeziehung jeder Aufgabengruppe noch einen Zugewinn an Reliabilität für den Testwert bringt. Durch Weglassen von „Schlauchfiguren“ oder „Figurenlernen“ wird die Prognosegüte des Studienerfolges für die erste Vorprüfung nicht wesentlich geringer. Diese beiden Aufgaben wurden aber im Test belassen, weil räumliches Denken bzw. räumliche Vorstellungen ggf. in späteren Studienabschnitten stärker gefordert werden.

Können Erfolgreiche wirklich drastisch häufiger zu einem Zulassungstest für das Medizinstudium angetreten sein? (S. 9)

Diese Aussage erweckt den Eindruck, dass es hier tatsächlich eine nennenswerte Häufigkeit grösser als 2 Teilnahmen gibt. In der Schweiz (EMS 1998 bis heute) werden praktisch keine EU-Bürger zum Studium und folglich zum EMS zugelassen. In Deutschland hat es den TMS letztmals 1996 gegeben – man durfte aber nur einmal am TMS überhaupt teilnehmen. 2007 im Mai hat es erstmals wieder einen TMS in Baden-Württemberg gegeben (auch jetzt darf wieder nur einmal am Test teilgenommen werden). Im Maximalfall könnten es also 2 vorherige Teilnahmen sein (2006 in Österreich, 2007 in Deutschland) – wobei es 2007 auch aus Österreich Testwiederholungen gibt. Es wäre möglich, dass die entsprechende Frage bei der Evaluation missverstanden wurde und sich beispielsweise auf alle Tests bezieht.

Die Empfehlung, eine veröffentlichte Originalversion unter Echtzeitbedingungen zu bearbeiten, wird wärmstens unterstützt. In den ersten Jahren des EMS waren Personen, die das zweite Mal am EMS teilnahmen, auch deutlich erfolgreicher als beim ersten Mal. Dieser Unterschied hat abgenommen, seit „Probeläufe“ mit der Originalversion empfohlen und unterstützt werden.

„Verrechnungsfairness“ und Bemerkungen zur Konstruktion des EMS (S.9)

Im Gutachten wird der Begriff „Verrechnungsfairness“ verwendet und es werden „Mängel“ dieser „Verrechnungsfairness“ – in beiden Tests – festgestellt. Betrachtet man die Definition, wo gefordert wird, „*dass Testaufgaben ausschliesslich die interessierende Kompetenz (bzw. den interessierenden Wissensbereich) erfassen sollen*“ (S.7), wird der Bezug auf die sogenannte probabilistische Testtheorie (PTT) bzw. Rasch-Skalierung deutlich, für welche dieses Konzept ausschliesslich gültig ist. Es ist **keinesfalls unumstritten,** dass Konzepte dieser Theorie hinsichtlich der Skalierung für alle Tests gelten müssen. **„Nicht verrechnungsfair“ nach Rasch bedeutet nicht zwingend, dass im „normalen“ Sprachsinne Fairness verletzt wäre.**

Der EMS wurde nach der klassischen Testtheorie konstruiert. Entsprechend sind deren Bewertungskonzepte und Gütekriterien anzuwenden und eigentlich ist jeder Streit müssig, weil ein valides prognoserelevantes Instrument wie nachgewiesenermassen der EMS auch alle anderen Gütekriterien erfüllen muss – denn sonst könnte es nicht valide sein.

Der „Streit“ ist allerdings alt und wurde/wird teilweise etwas „fundamentalistisch“ geführt. Trotzdem es die probabilistische Testtheorie schon seit 1960 gibt, sind die wichtigsten und anwendungshäufigsten Leistungstests der Psychologie nach wie vor auf der Basis der Klassischen Testtheorie entwickelt. Auch wichtige amerikanische Studierfähigkeitstests, die seit mehr als einem halben Jahrhundert zum Teil weltweit und mit Teilnehmerzahlen in Millionenhöhe verwendet und kontinuierlich von ausgewiesenen Testexperten weiterentwickelt werden, fussen bis heute auf der Klassischen Testtheorie. Beide Strategien haben Vor- und Nachteile, beide können zu seriösen Tests führen. Für nach PTT konstruierte Tests muss man eine „latente Eigenschaft“ (Dimension) annehmen und durch eine Skala messen. Die Skala muss vergleichsweise restriktiven messtheoretischen Voraussetzungen folgen, was praktisch auch heisst: homogen bezüglich der Aufgaben sein, um „eindimensional“ *die interessierende Kompetenz* zu messen. Das erreicht man praktisch nur, indem alle Aufgaben auch relativ gleichartig konstruiert sind.

Die Frage, ob man solche Tests verwenden soll, wurde bei der Konstruktion des TMS im damaligen „Beirat für psychologische und allgemeine Fragen der Testentwicklung und des Testeinsatzes bei der Hochschulzulassung“, dem die seinerzeitige „Crème“ der Psychologie in Deutschland angehörte, ebenfalls diskutiert. Man hat sich bewusst für einen anderen Weg entschieden, komplexere und für ein Studium anforderungsnähere Tests mit mehr inhaltlicher Validität zu verwenden. Der Beirat empfahl mit Nachdruck, der TMS solle nicht homogene, **sondern komplexe Aufgaben enthalten und damit der Realität eines Studiums, in dem ganz verschiedene kognitive Funktionen gleichzeitig gefordert sind, möglichst nahe kommen**. Auf diese Weise lasse sich dank der Simulation solcher komplexer Anforderungssituationen eine höhere inhaltliche Validität erzielen. Eine Studienanforderung wie das Lesen von Tabellen, das Verstehen von Texten muss nicht psychologisch „eindimensional“, sondern kann wie im realen Leben komplex sein. Auch gibt es verschiedene Arten von Diagrammen oder Texten, die innerhalb der Aufgabengruppen variiert werden, um das vorherige Üben einzelner Prototypen zu erschweren.

Damit wird auch klar, warum die **Trennschärfen** der Aufgabengruppen tendenziell geringer sind als bei homogeneren Tests, wo das Ergebnis dieser Aufgabengruppe dann einzeln diagnostisch interpretiert werden soll. Eine Optimierung der Aufgabengruppe auf hohe Trennschärfe wie im Gutachten gefordert hätte die gewollte Aufgabenvielfalt verringert. In der jetzigen Höhe sind sie ein Optimum zwischen Aufgabenvielfalt und Sicherstellung, dass die richtige Lösung von den Leistungsbesten mit grösserer Wahrscheinlichkeit gewählt wird.

Dass die geprüfte Rasch-Skalierung nicht so eindeutig die „Verrechnungsfairness“ bzw. die Unterscheidung von „guten“ und „schlechten“ Personen aufklärt, zeigt eine Gegenüberstellung des tatsächlichen Beitrages der einzelnen Aufgabengruppen zur Unterscheidung der Personen nach

Prüfungserfolg[5] und vermeintlichen „Problemen mit der „Verrechnungsfairness“. Die Prognosekraft (Validität) ist dabei ein externes Kriterium und entscheidet damit „objektiv“ über die Testgüte: Je höher das Gewicht in einer standardisierten Diskriminanzfunktion zur Trennung der Gruppen nach dem Bestehen der Prüfung, umso „wertvoller“ ist eine Aufgabengruppe. Die Aufgabengruppen mit dem höchsten Prognosebeitrag sind auch diejenigen, welche bezogen auf die Anforderungen „mehrdimensional“ und komplex sind.

Angeblich hätten nun diese für den Test aussagefähigsten Untertests die grössten „Probleme“. Da die Praxis das Kriterium der Wahrheit ist, sollte man die Definition von „Problem“ noch einmal überdenken. Wir halten es im Übrigen nicht für möglich, im Rahmen dieses Projektes den fachlichen „Streit“ Klassische vs. Probabilistische Testtheorie zu entscheiden – dies ist auch sonst kaum gelungen.

	Gewicht in einer standardisierten Diskriminanzfunktion zur Trennung der Gruppen nach Prüfungserfolg (je höher, desto valider ist der Untertest für die Erfolgsprognose)	Angebliche “Probleme” lt. Gutachten
Quantitative und formale Probleme	.695	X
Diagramme und Tabellen	.665	
Med.-naturwiss. Grundverständnis	.644	X
Textverständnis	.612	X
Muster zuordnen	.489	X
Konzentr. und sorgf. Arbeiten	.426	
Schlauchfiguren	.382	
Figuren lernen	.361	
Fakten lernen	.352	

Misst der Test das Gleiche für Frauen und Männer, für Deutsche und Österreicher? (S.9)

Es wird behauptet, dass eine Reihe von Aufgaben und Untertests bei Frauen und Männern sowie bei Deutschen und Österreichern nicht das gleiche messen würde. Auch diese Annahme beruht letztendlich darauf, dass die Rasch-Skalierung für die gemessenen Merkmale angemessen wäre. Wir vermuten, dass dies ein Problem der Parameterschätzung im Rasch-Modell ist – weil es zwischen diesen Gruppen objektiv vergleichsweise grosse Verteilungsunterschiede bei den Ergebnissen in einzelnen Aufgabengruppen gibt.

Weil Fairness in der Schweiz wegen der Anwendung des EMS in drei Sprachgruppen besonders wichtig ist, wurde mit den Analysen zum Differential Item Functioning (DIF) eine international eingeführte Methodik zur Identifikation und Korrektur solcher Effekte auch hier eingesetzt. Sie beruht auf der Annahme, dass Items, die Unterschiedliches messen, in einer Gegenüberstellung der Item-Schwierigkeiten für jeweils zwei betrachtete Gruppen sich von einem allgemeinen Trend über alle Items unterscheiden (Delta-Plot). Die Details werden im Bericht 2007 des EMS-AT bzw. allgemein z.B. bei Hänsgen und Spicher (2007) dargestellt. Es zeigen sich sowohl in der Gegenüberstellung Männer zu Frauen, als auch in der Gegenüberstellung Österreich zu EU keine Hinweise, dass solche Unterschiede eine Bedeutung haben, Wären solche

[5] Vergleiche Bericht 7 des ZTD (2001). Je höher der Wert, desto höher ist das Gewicht eines Untertests für die Trennung von Erfolgreichen und nicht Erfolgreichen in einer Diskriminanzfunktion. Verwendet wurden die Daten der Studienanfänger 1998 und 1999 und das Bestehen der ersten Vorprüfung.

Effekte nachgewiesen worden, wäre auch ein Grund für einen Ausgleich z.B. in Form von Bonuspunkten vorhanden gewesen.

Mittels Faktorenanalyse wurde des Weiteren nachgewiesen, dass der TMS, der EMS in der Schweiz und der EMS in Österreich jeweils die gleiche Faktorenstruktur (sogar ohne Ähnlichkeitsrotation) aufweisen – was ein weiterer Hinweis auf Vergleichbarkeit darstellt.

Ist der EMS auf die Studienanforderungen in Österreich bezogen? (S. 11)

Im November 2005 sind beide Medizinuniversitäten sehr kurzfristig an die Rektorenkonferenz der Schweizer Universitäten (CRUS) herangetreten, ob die notwendige Beschränkung der Zulassung ab 2006 auf der Basis des Eignungstests für das Medizinstudium (EMS) erfolgen kann. Es gab zwei wesentliche Argumente für alle Verantwortlichen, den EMS auch in Österreich einzusetzen:

- Der Test wurde in Deutschland und der Schweiz umfangreich positiv evaluiert. Die prognostische Validität war in beiden Ländern sehr gut, auch andere Anforderungen für ein solches Verfahren (elaboriertes Informationssystem mit Test Info und im WEB, ausreichende Vorbereitungsmöglichkeiten stehen offiziell zur Verfügung) wurden erfüllt.
- Ein erster Abgleich der Studienanforderungen zeigte, dass diejenigen von Österreich nicht wesentlich von denen in Deutschland und der Schweiz abweichen. Die seinerzeit in Deutschland durchgeführten Anforderungsanalysen für ein Studium der Medizin wurden sehr aufwändig und von verschiedenen Seiten durchgeführt (siehe Trost, 1998). Auf der Grundlage dieser Analysen wurden ca. 50 Bereiche und Aufgabentypen geprüft und die unter vielen Aspekten am besten geeignet erscheinenden 13 für die Erprobungen im Übergangsverfahren des TMS ausgewählt und empirisch erprobt. Von diesen haben vor allem aufgrund der Ergebnisse zur Validität dann 9 Aufgabengruppen Eingang in den TMS gefunden, eine (Planen und Organisieren) kam dann im EMS dazu.

TMS und EMS sind nicht statisch, die Übereinstimmung mit den Studienanforderungen wird fortlaufend beobachtet (u.a. Workshops mit den Fakultätsvertretern), Veränderungen und Erweiterungen werden vorgenommen. Richtig ist, dass sich die Medizinuniversitäten aus Österreich in diesen Prozess einbringen sollten, wenn der EMS dort weiter eingesetzt werden soll. Dies beinhaltet sowohl eine genauere Analyse der Passfähigkeit der EMS-Anforderungen für die Studienanforderungen in Österreich, als auch das Einbringen neuer Anforderungen in die geplante Weiterentwicklung des EMS.

Wie ist die „Bezweifelung" der Prognosekraft des TMS/EMS zu werten, wenn es Änderungen der Studienorganisation, der Curricla, der Lehr- und Lernformen sowie des Prüfungsmodus gäbe (S. 12)

Hier wird eine notwendige Einschränkung missverstanden, Bei TMS und EMS wird tatsächlich **immer** angemerkt, dass es eine kontinuierliche Passung zwischen Test (als dem Vorhersageinstrument) und Studienerfolg (als dem Vorherzusagenden) gibt, der aktuelle Test erst einmal nur auf die Vorhersage des aktuellen Studienerfolg ausgerichtet ist. Ändern sich die Erfolgskriterien, muss sich auch der Test ändern. Diese Entwicklung gab und gibt es aber. Schon in Deutschland wurde der Test mehrfach angepasst, es gab bekanntlich eine lange Evaluationsphase. In der Schweiz wurde der Test ebenfalls weiter angepasst – ein Untertest Planen und Organisieren ergänzt und einer (Konzentriertes und Sorgfältiges Arbeiten) modifiziert. Weitere Veränderungen sind geplant, auch hinsichtlich der Erfassung sozialer Fähigkeiten.

In der Schweiz wurden die Studienreformen (problemorientierter Unterricht, verbesserte Betreuungsverhältnisse) ebenfalls umgesetzt und eine neueste Evaluation zeigt, dass auch unter diesen Bedingungen der aktuelle EMS den Studienerfolg sehr gut vorhersagt [6].

[6] http://www.unifr.ch/ztd/ems/emseval07.pdf

Soziale Kompetenzen berücksichtigen? (S11)

Es wird richtig festgestellt, dass die Erfassung sozialer Kompetenzen (kommunikative und sozialkognitive Kompetenzen) im EMS noch nicht erfolgt. Eine Ursache ist, dass eine testpsychologische Erfassung bisher problematisch ist (vgl. Hänsgen 2007)[7]. Die Bedeutung solcher Fähigkeiten steigt mit veränderten Studienanforderungen und es ist eine der laufenden Entwicklungsanstrengungen, zumindest die Fähigkeiten auf diesem Gebiet zu erfassen (die aufgrund von Untersuchungen in Belgien[8] aber offenbar aufwändige Videopräsentationen voraussetzen).

Es wäre auch zu diskutieren, ob in Deutschland dies durch die Berücksichtigung anderer Kriterien wirklich „aufgefangen" wurde, wie im Gutachten behauptet. Es gab zu Zeiten des TMS verschiedene Zulassungsquoten (siehe Trost 1994, S. 139):

- ca. 10% der Plätze gingen als „Vorab-Quote an ausländische Bewerber, Härtefälle, Zweitstudienbewerber und Bewerber mit „besonderer Hochschulzugangsberechtigung" oder „bevorzugt Zuzulassende" (z.B. früher wegen Militärdienst zurückgestellt).
- Von den restlichen Plätzen wurden in der folgenden Reihenfolge vier Quoten bedient. Die Zahl der Bewerbungen um einen Studienplatz überstieg die Kapazität zu dieser Zeit um das Acht- bis Neunfache.
    - 45% nach einer Kombination Abiturnote (mit Ausgleich der Länderunterschiede) und TMS Die Abiturnote allein benachteiligte, wenn man Schul- und Studienleistung in Beziehung setzt, auch dort die Männer wegen einer offenbar strengeren Benotung (siehe Trost 1996).
    - 10% nach dem Testergebnis im TMS
    - 20% nach der Wartezeit (Zahl der Bewerbungssemester), die Hartnäckigkeit belohnte, aber sicher auch von den finanziellen Möglichkeiten beeinflusst war, eine Überbrückungszeit mit einer anderen Ausbildung oder Tätigkeit einzulegen. Hier spielten Leistungskriterien dann keine Rolle und Beharrlichkeit ist nur ein marginaler Aspekt sozialer Kompetenzen.
    - 15% wurden aufgrund eines Interviews von den Universitäten ausgewählt, wobei maximal die dreifache Menge an Personen per Los (!) entsprechend der gewünschten Studienorte den Universitäten zugeteilt wurde. Es sind die niedrigen Prognosewerte für Studienerfolg von Interviews zu berücksichtigen, die in der Metaanalyse von Schuler lediglich bei Korrelationen um 0.10 bis 0.20 liegen.

Die Angehörigen der Wartezeit- und Auswahlgesprächsquote traten später zur ärztlichen Vorprüfung an und hatten – unabhängig vom Zeitpunkt – auch die geringsten Erfolgsraten (Trost 1994, S. 142 f). Die Leistungen dieser beiden Gruppen waren in schriftlichen und mündlichen Prüfungen verglichen mit den anderen Quoten am schlechtesten (S. 150).

Als die Schweiz ein eigenes Zulassungsverfahren konzipierte, wollte man auch aufgrund dieser Erfahrungen weder Interview noch Wartezeit berücksichtigen. Wegen der unterschiedlichen

---

[7] Die meisten in standardisierten Tests verwendbaren Methoden beruhen auf Selbsteinschätzungen des eigenen Erlebens und Verhaltens. Dies kann im Sinne der sozialen Erwünschtheit leicht beeinflusst werden (ggf. durch Trainingskurse erlernt) – deshalb ist die Verwendung problematisch. Für Interviews ist auch bekannt, dass ein optimales Verhalten trainiert werden kann. Die Prognosekraft von Interviews ist in der Metaanalyse von Schuler im Übrigen enttäuschend gering.

[8] Lievens verwendet eine Videopräsentation sozialer Situationen, die einzuschätzen sind. Die Prognosekraft für Studienerfolg ist insgesamt gering. Lievens und Sackett (2007) weisen zudem nach, dass eine Präsentation auf Papier noch geringere Prognosewerte aufweist. In Belgien wird der Test an einem Ort in einem Saal durchgeführt. Die Gewährleistung gleichguter Präsentationsbedingungen in den verschiedenen Testorten und Testlokalen wäre in der Schweiz und in Österreich zumindest extrem aufwändig und der erwartete Nutzen ist bisher wegen der geringen Prognosekraft im Verhältnis dazu zu gering.

Notenmassstäbe in den (teilweise recht kleinen) Kantonen schien es auch praktisch unmöglich, die Maturitätsnoten zu berücksichtigen. Vor allem fürchtete man eine negative Rückwirkung auf die schulische Praxis der Notengebung, wenn von der Note die Zulassung zum Studium abhängt. Dies würde überall dort eine Rolle spielen, wo es keine zentralen und standardisierten Abiturprüfungen gibt.

Sonstige Anmerkungen

Die praktischen Empfehlungen des Gutachtens hinsichtlich des EMS können nur bekräftigt werden:

- Betonung der Vorbereitung anhand der offiziell empfohlenen Vorbereitungsstrategie und Unterstützung durch Schulen, weil gemeinsame Vorbereitung einer Vorbereitung allein auch überlegen ist. Dabei sollte eine Bearbeitung der veröffentlichten Originalversion unter Echtzeitbedingungen erfolgen.
- Einbringen der österreichischen Wünsche, Bedürfnisse und Erfahrungen in den Prozess der kontinuierlichen Weiterentwicklung des EMS, wenn eine langfristigere Zusammenarbeit gewünscht ist.

Ergänzend würden wir vorschlagen, sich auch mit „**Neigungen**“ und der Entwicklung von Informations- und „Selbstberatungsmöglichkeiten“ zu beschäftigen (vgl. dazu Hänsgen 2007). Neigung wäre alles, was mit Interesse und Motivation für ein Studienfach umschrieben werden kann. Bei fehlender Neigung wird ein Studium sicher gar nicht aufgenommen. Problematischer sind die Fälle, wo man während des Studiums entdeckt, dass ein Fach nicht den eigenen Neigungen entspricht. Dies kann an fehlenden oder falschen Informationen über Studium und Beruf liegen. Die Erfassung von Neigungen und deren Verwendung als Zulassungskriterium ist aber wegen der hohen Anfälligkeit für Trainierbarkeit und Verfälschung („sich im besten Licht darzustellen“) praktisch unmöglich.

Eine in anderen Ländern erfolgreich eingesetzte Methode sind **Studienberatungsinstrumente**, die zur „Selbstberatung“ angeboten werden und meist im Internet realisiert sind. Es werden detaillierte Informationen über die verschiedensten Anforderungen von Studium und späterem Beruf gegeben und Checklisten sowie Tests angeboten, um ohne Druck eigene Fähigkeiten (hier auch soziale Kompetenzen) und Neigungen zu erkunden. Da hier das Interesse der Personen an realistischen Aussagen überwiegt - weil es nicht als Zulassungskriterium verwendet wird - ist es auch nicht sinnvoll, sich in einem „besseren Licht“ darzustellen. Es gibt mittlerweile Universitäten, welche die **Absolvierung** eines solchen Studienberatungstestes zur Pflicht machen, ohne die Ergebnisse selbst für die Zulassung zu verwenden. Es wäre eine Herausforderung mit grossem erwarteten Nutzen, für Medizin ein solches Selbstberatungsinstrument zu entwickeln – die meisten uns bekannten Instrumente beziehen sich bisher auf wirtschaftliche oder technische Studiengänge.

Perspektive

Einen für die Zulassung geeigneten Test zu entwickeln und laufend an die sich verändernden Studienbedingungen anzupassen, ist eine grosse Herausforderung an die dafür notwendigen Ressourcen – dies wird durch das Gutachten auch deutlich. Ein Land wie die Schweiz wäre aus Kostengründen gar nicht allein in der Lage gewesen, ein solches Projekt wie den EMS durchzuführen. Auch in Deutschland wurde der TMS seinerzeit mit aus Kostengründen eingestellt (und man hat ihn heute reaktiviert, weil ein faires und funktionierendes Zulassungskriterium nur auf der Basis der Abiturnoten und der anderen Kriterien suboptimal ist).

Es gäbe eine Chance, dass sich alle drei Länder (Deutschland, Schweiz und Österreich) „paritätisch“ so zusammentun, dass die weitere Testentwicklung gemeinsam erfolgt. Auf der Basis eines allgemeinen Rahmenkonzeptes kann man den Spezifika der Anforderungen der Länder bei Testkonzeption und Weiterentwicklung ausreichend gerecht werden (was die entsprechende Zusammenarbeit Deutschland-Schweiz belegt). Vergleicht man dies mit den Ressourcen ent-

sprechender Institute im Ausland (Educational Testing Service bzw. AAMC in den USA, welche den dortigen Mediziner-Test betreuen), würde sich dann auch im deutschen Sprachraum eine entsprechende „kritische Masse“ ergeben, um jederzeit eine moderne Testentwicklung zu garantieren (auch mit der entsprechenden Grundlagenforschung) und die Kosten für alle Seiten zu minimieren. Dafür wären Strukturen einer paritätischen Zusammenarbeit zu finden, eine grundsätzliche Bereitschaft ist vorhanden.

Literatur:

Frischenschlager O., Mitterauer L., Haidinger G (2005): Leistungsfaktoren als potenzielle Auswahlkriterien im Medizinstudium. E-ZfHD und Zeitschrift für Hochschuldidaktik, Heft 6, Dezember 2005.

Hänsgen, K.-D., Spicher, B. (2001): EMS • Eignungstest für das Medizinstudium in der Schweiz. Vorhersage des Prüfungserfolges. Bericht 7. Fribourg: ZTD http://www.unifr.ch/ztd/ems/berichte/b7/ztd_bericht_7_EVA.pdf

Hänsgen, K.-D (2007): Numerus clausus in der Medizin – werden die Richtigen ausgewählt für Studium und Beruf? Schweizerische Ärztezeitung 2007;88: 46 http://www.unifr.ch/ztd/ems/berichte/2007-46-1078.pdf

Hänsgen, K.-D, Spicher, B. (2007). EMS Eignungstest für das Medizinstudium in der Schweiz Bericht 13 über die Durchführung und Ergebnisse. Fribourg: ZTD http://www.unifr.ch/ztd/ems/berichte/Bericht13.pdf

Trost, G. (Hrsg.) (1994). Test für Medizinische Studiengänge (TMS): Studien zur Evaluation (18. Arbeitsbericht). Bonn: ITB.

Trost, G. (1996): Testergebnisse versus Schulnoten als Auswahlkriterien: Paternoster-Effekt, Filter-Effekt, Kosten-Nutzen-Effekte und Auswirkungen auf die Fairneß der Zulassung. In: Hänsgen u.a. (Hrsg) (1996): Bericht 2 des Zentrums für Testentwicklung.

Trost, G. u.a.(1998). Evaluation des Tests für Medizinische Studiengänge (TMS): Synopse der Ergebnisse. Bonn: ITB.

Prof. Dr. Klaus-Dieter Hänsgen

Direktor des Zentrums für Testentwicklung

ZTD Zentrum für Testentwicklung und Diagnostik

Universität Fribourg

UNI Rte Englisberg 9

CH-1763 GRANGES-PACCOT

Phone : +41 <0>26 300 7989/86

Fax: +41 <0>26 300 9763

Email : Klaus-Dieter.Haensgen@unifr.ch

WEB : http://www.unifr.ch/ztd/

Anhang

Nachfolgend seien alle sachlichen Argumente bezüglich der Genderfairness tabellarisch zusammengetragen, die von uns bisher betrachtet wurden. Wir unterscheiden dabei vier mögliche Ursachenkomplexe für die Unterschiede zwischen den Geschlechtern beim EMS. Man kann jeweils bestimmen, ob sich durch die dargestellten Fakten Ursachen ausschliessen lassen oder nicht. Es ist markiert, ob der jeweilige Fakt eher für (+) oder gegen (-) eine der Hypothesen spricht:

- **TEST:** Allein testbedingte Unterschiede, die keine Entsprechung im vorherzusagenden Bereich haben: Durch den Test selber würden einzelne Gruppen benachteiligt. Es gibt bisher kein einziges Faktum, was für wirklich testbedingte Ursachen spricht – wenn man akzeptiert, dass der Test hinsichtlich der Vorhersage des Studienerfolges so sein muss wie er ist. Letzteres ist mittlerweile durch die Evaluationsbefunde aus Wien auch belegbar.

- **Bedingungen:** Ursache sind die Abnahmebedingungen, z.B erstmals einen derartigen Test, grosse Räume mit vielen Personen, strenge Aufsicht und Kontrollen. 2007 wurde grosser Wert darauf gelegt, die Bedingungen stressfreier zu gestalten - trotzdem bleibt der Unterschied vorhanden. Wenn man von einer unterschiedlichen Ansprechbarkeit der Geschlechter auf Stress ausgeht, kann die Bedeutung der Abnahmebedingungen nicht ganz ausgeschlossen werden – es ist aber wohl nicht der Hauptfaktor.

- **Studienwahl:** Es handelt sich nicht um repräsentative Stichproben für Männer und Frauen, sondern um Medizinstudiumsbewerbungen. Hier können sich aus beiden Gruppen Personen mit unterschiedlichem Leistungsniveau bewerben. Wenn z.B. Männer strenger benotet werden, könnten leistungsschlechtere Männer vor einer Bewerbung für Medizin eher zurückschrecken (weil es als anspruchsvolles Studium bekannt ist). Würden Frauen besser benotet als es dem realen Leistungsniveau entspricht, würden diese eher den Mut finden, sich für Medizin zu bewerben. Insbesondere die PISA-Befunde sprechen für diesen Faktor. In repräsentativen Stichproben der 15jährigen finden sich diese Leistungsunterschiede offenbar nicht so wie beim EMS.

- **Aktueller wahrer Unterschied:** Es kann einen tatsächlichen Leistungsunterschied zwischen beiden Geschlechtern geben. Der muss nicht unabänderlich sein, sondern kann auch einen aktuellen Unterschied betreffen, der sich noch ausgleichen lässt. Sollte die unterschiedliche naturwissenschaftliche Orientierung, die unterschiedliche Beschäftigung mit dieser Materie in verschiedenen Schulen wie diskutiert hier eine Rolle spielen, würde dies hierzu gerechnet – ohne damit zu unterstellen, dies wäre durch mehr Beschäftigung nicht ausgleichbar. Auch für diesen Fall finden sich Belege.

Offensichtlich wirken mehrere Ursachen zusammen, was die Aufklärung erschwert. Alle Fakten sprechen allerdings klar dagegen, dass es sich um ein Problem des EMS handelt.

	Test	Bedin-gungen	Studien wahl	Wahrer Unter-schied
Ländervergleich **Schweiz**: gleicher Tag, gleicher Test: keine entsprechenden Unterschiede zwischen Geschlechtern beim Test 2006 und 2007	-	+	+	+
Ländervergleich **Deutschland**: gleiche Testaufgaben: keine entsprechenden Unterschiede in allen Sessionen. Auch 2007 in Baden-Württemberg keine derart grossen Unterschiede	-	+	+	+
2006/07 Gleiche Ergebnisse mit Test eines ganz anderen Typs (Wissenstest) in **Graz** – gleiche Auswahlsituation, auch dort die Leistung der Männer besser	-	+	+	+
DIF-Analyse („Differential Item Functionning“) auf Itemebene: kein Itembias 2006 und 2007, d.h. es gibt keine einzelnen Aufgaben die z.B. aufgrund unterschiedlicher Erfahrungen der Geschlechter von einer Gruppe speziell schlechter beantwortet werden	-	+	+	+
Unterschiede Studienerfolg für Männer Frauen MedUni Wien (Studien Frischenschlager u.a) in 2 Jahrgängen Nachweis, dass Männer auch bessere Prüfungsleistungen erreichen, was der Test dann korrekt vorher sagt.	-	-	+	+
Evaluation Wien: gute Vorhersageleistung SIP1 durch EMS für Kohorte 2006, in Korrelation kein Unterschied Männer zu Frauen, gleicher Mittelwert für die Gruppen nach Nichtbestehen der Prüfung für die Geschlechter.	-	-	+	+
Befunde Linz (Brandstätter) Studienberatung und Auswahl, spezielle Disziplinen: Männer sind auch dort besser als Frauen	-	-	+	+
Befunde Salzburg (Lengenfelder, Baumann), Studienauswahl Salzburg: Männer ebenfalls besser.	-	-	+	+
PISA-Studie 03 und 07: Mathematik, Naturwissenschaften und Problemlösen im 15. Jahr bei repräsentativen Stichproben nicht der gleiche Unterschied wie im EMS.	-	-	+	-
Jungen werden bei der Benotung **benachteiligt**, strenger benotet (Studie Eder, Salzburg), daher weniger Mut zur Bewerbung Medizin bei gleicher Leistung wie Mädchen?	-	-	+	-
Stressituation Test in AT grösser? 2006 Unterschiede am Vormittag grösser, geringer, wenn Test leicht erlebt wurde, wenn Tagesform besser - grösser, wenn weniger Vorbereitungszeit und in **kleineren** Testlokalen (Innsbruck). 2007 wurde die Situation allerdings deutlich stressfreier gestaltet, der Effekt bleibt trotzdem.	-	(+)	-	-