Zentrum für Testentwicklung und Diagnostik
am Psychologischen Institut der Universität Fribourg

Der Eignungstest für das Medizinstudium in der Schweiz

Ein Probelauf

Rainer Hofer, Daniel Ruefli & Klaus-Dieter Hänsgen

Bericht 1 (1995)

Inhalt

Zentrum für Testentwicklung und Diagnostik
am Psychologischen Institut der Universität Fribourg
Route d'Englisberg 9, CH-1763 Granges-Paccot
Tel. 037/ 29 79 86 - Fax 037/ 29 97 63 - e-mail: ZTD@UNIFR.CH
Internet: http://www.unifr.ch/Main/ztd.html

Zusammenfassung

Die Einführung eines Numerus Clausus (NC) für das Medizinstudium in der Schweiz ist eine politische Entscheidung, die unter Abwägung verschiedener Werte getroffen werden muss. Ein NC kann allerdings nur umgesetzt werden, wenn ein anerkanntes, zuverlässiges und möglichst gerechtes Kriterium zur Verfügung steht, um diese Zulassung zu regeln. Bei der weiteren Auseinandersetzung mit diesem Thema sollte genauer als bisher berücksichtigt werden, dass

- die politische Entscheidung die Notwendigkeit der Begrenzung mittels NC und die Studienplatz-Kapazität vorgibt;
- das Kriterium ***unter der Bedingung dieses NC*** danach zu bewerten ist, ob es ein relativ faires bzw. das fairste ist.

Eine gesamtschweizerisch koordinierte Zulassung zum Medizinstudium wird nach wie vor angestrebt. Bisher lagen noch keine Erfahrungen vor, ob ein Eignungstest in der Schweiz in ähnlicher Weise wie in Deutschland durchführbar ist. Beispielsweise muss der Test hier im Unterschied zu Deutschland in drei Landessprachen absolviert werden können. Die Gewährleistung der unabdingbaren Chancengleichheit war deshalb in den bisherigen Diskussionen zum Test immer ein Schwerpunkt und es wurden nicht wenige Befürchtungen diesbezüglich geäussert.

Ein Probelauf des Tests in der Schweiz diente deshalb vor allem der Klärung folgender Fragen:

- Ist die **Differenzierungsfähigkeit** nach der Leistung als Zulassungskriterium überhaupt geeignet? Oder ist der Test allgemein zu leicht oder zu schwierig, so dass es keine Differenzierungsmöglichkeiten auf der Basis der Testleistungen gibt?
- Sind die **Sprachformen** der Tests äquivalent und kann in allen Sprachgruppen von einer gleichen Zuverlässigkeit und Fairness ausgegangen werden?
- Sind Unterschiede in Testleistungen zwischen beiden **Geschlechtern** oder zwischen verschiedenen **Maturitätstypen** derart vorhanden, dass eine Chancengleichheit möglicherweise nicht gegeben ist?

Zur Überprüfung der psychologischen Gütekriterien und der Fairness des Tests wurde im Collège Sainte-Croix in Freiburg (Schweiz) mit einer Stichprobe von 54 deutschsprachigen und 126 französischsprachigen Gymnasiastinnen und Gymnasiasten ein Probelauf durchgeführt und dessen

Resultate mit denen der Testdurchführung des Jahres 1992 in Deutschland (als Kontrollgruppe) verglichen.

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.

Eine wichtige Frage ist nach wie vor das Verhältnis der Geschlechter bezüglich der Testleistungen. Die aus den deutschen Ergebnissen abgeleitete Hypothese, dass **Frauen tendenziell schlechtere** Leistungen erreichen als Männer, konnte **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.

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 Absolventen 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 eindeutig, 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.

Problemstellung

Die Einführung eines Numerus Clausus erfordert ein objektives, anerkanntes Zulassungskriterium, nach welchem dieser Zugang geregelt wird. Dieses Kriterium muss gerecht sein, die Gleichbehandlung aller Sprachgruppen und beider Geschlechter etc. gewährleisten. Als Kriterium für die geplante Zulassungsbeschränkung zum Medizinstudium in der Schweiz ist ein Eignungstest vorgesehen. Dieser Eignungstest, der am ehesten als ein „Probestudium“ mit typischen Aufgabenstellungen für das Medizinstudium zu beschreiben ist, wurde in Deutschland (Trost et al. 1977-1994) entwickelt und an die Schweizer Verhältnisse adaptiert. Seine Ergebnisse erlauben eine Aussage über die Studierfähigkeit der einzelnen Testpersonen. Bezüglich der Vorhersagekraft für den Studienerfolg gehört dieser Test im internationalen Vergleich zu den besten Verfahren.

Bei den Diskussionen um den Test in der Schweiz stützte man sich auf die deutschen Ergebnisse, da dort der Test seit 1986 eingesetzt wird und seither in zahlreichen wissenschaftlichen Studien den Beweis seiner Eignung erbracht hat. Diese deutschen Ergebnisse hatten jedoch zu Behauptungen geführt, der Test sei nicht spezifisch genug für das Medizinstudium oder er diskriminiere die Frauen und diejenigen, welche zu “sozialen Unterschichten“ gehören (Hänsgen et al. 1995).

Es soll in dieser Studie überprüft werden, ob Behauptungen dieser Art zutreffen oder ob man den Test “geprügelt“ und den Numerus Clausus gemeint hat. Da der deutsche Test für die Schweiz nicht nur redaktionell überarbeitet, sondern auch in die Landessprachen Französisch und Italienisch übersetzt wurde, müssen dessen Ergebnisse hier nicht genauso zutreffen.

Folgende Fragen wurden mit Hilfe des Test-Probelaufs in Freiburg überprüft:

- Kann der Eignungstest in der Schweiz unter vergleichbaren Bedingungen wie in Deutschland durchgeführt werden?
- Wie hoch sind die Gütekennwerte des Eignungstests im Probelauf? Zeigen sich Unterschiede im Vergleich zu einer deutschen Kontrollgruppe?
- Zeigen sich Geschlechterunterschiede auf Test-, Untertest- oder Itemebene?
- Zeigen sich infolge der Übersetzung des Tests in das Französische kulturelle oder sprachliche Unterschiede auf Test-, Untertest- oder Itemebene?

- Wie wirken sich Faktoren wie Alter, Maturitätstyp, Interessensschwerpunkte in der Schule, Schulnoten oder Motivation auf die Testergebnisse aus?

Methode

Stichprobe

Die Untersuchung fand im Collège Sainte-Croix in Freiburg statt. Es wurden Schülerinnen und Schüler gewonnen, die ein Jahr vor dem Maturitätsabschluss standen.

Im Vergleich zu Deutschland fehlte natürlich bei dieser Stichprobe eine Bewerbungsmotivation. Dazu kam, dass der Test-Probelauf den Schulbetrieb sowenig wie möglich stören durfte, weshalb seine Durchführung auf Ende des Schuljahres, also nach den Prüfungswochen, geplant wurde. Zur Förderung der intrinsischen Motivation konnte letztendlich nebst dem Erfahrungsgewinn das Angebot einer relevanten Leistungsrückmeldung gemacht werden.

Instrumente und Durchführung

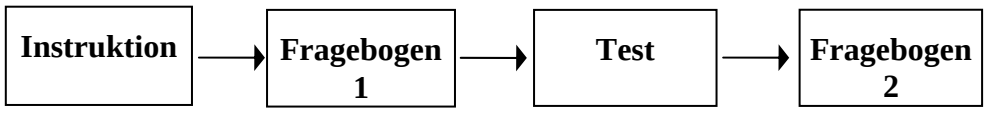
Verwendet wurden Aufgaben eines veröffentlichten Originaltests des Tests für Medizinische Studiengänge (TMS), der in deutscher Sprache vorliegt (Institut für Test- und Begabungsforschung 1990). Für eine Publikation als Vorbereitungsmaterial auf einen möglichen Test wurde er in die französische (und italienische) Sprache übersetzt.

Seine Ergebnisse stehen eindeutig und nachweislich mit der Studieneignung in Zusammenhang, denn diese wird als Kriterium der Testentwicklung mit verwendet. Die Anforderungen des Tests sind studiennah im Sinne eines „Probestudiums“, weil bei der Aufgabenerstellung Medizindozenten und Gymnasiallehrer mit am Tisch sitzen und die Aufgaben dahingehend entwickeln bzw. beurteilen, ob sie repräsentativ für die Studienanforderungen sind. Der Test ist keine Wissensprüfung, sondern es wird die Fähigkeit zum Denken und Problemlösen geprüft. Spezielles Wissen der Medizin wird nicht vorausgesetzt, sondern die für die Lösung notwendigen Fakten werden in der Aufgabe selbst mitgeteilt. Das sollte deshalb eine für das Studium sehr typische Anforderung sein.

Die Anwendung des Tests ist insofern gerecht, dass die Chancengleichheit durch standardisierte Durchführungs- und Auswertebedingungen gewährleistet ist. Die „Willkür“ eines Urteilers kann nicht zum Tragen kommen.

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

Das Vorgehen im Test-Probelauf ist im folgenden Design festgehalten:



INSTRUKTION

Zur Vorbereitung auf den Test gehört, sich vorab mit den Aufgabentypen, den Instruktionen und dem Testablauf vertraut zu machen. Eine Woche vor dem Test-Probelauf wurden deshalb die Gymnasiastinnen und Gymnasiasten über den Ablauf der Testung informiert. Die verschiedenen Aufgabentypen wurden dabei besprochen. Es wurde auch ein TEST-INFO (Zentrum für Testentwicklung und Diagnostik 1995) als Vorbereitungsmaterial bereitgestellt, um die Möglichkeit zu bieten, sich zu Hause auf den Test vorbereiten zu können. Dieses TEST-INFO entspricht dem Material, welches auch bei einer tatsächlichen Testanwendung zum Einsatz käme und für den geplanten Testeinsatz 1995 vorbereitet worden war.

Die insgesamt geringe Trainierbarkeit des Tests ist bekanntlich darauf zurückzuführen, dass der Test kein Faktenwissen prüft sondern Problemlösungen verlangt, die allein aus den Angaben der Aufgaben (Hänsgen et al. 1995) gewonnen werden können. Das vorherige Vertrautmachen mit den Anforderungen ist vor allem wichtig, um während der Testung keine wesentliche Zeit mehr für das Verstehen der Instruktionen zu benötigen und sofort mit der Aufgabenbearbeitung beginnen zu können.

FRAGEBOGEN 1

Um verschiedene Angaben über soziale und edukative Merkmale der Testpersonen zu erhalten, wurde anlässlich der Instruktion ein Fragebogen abgegeben. Es war den Gymnasiastinnen und Gymnasiasten freigestellt, die Fragen nach dem Alter, den Interessensschwerpunkten in der Schule, den Durchschnittsnoten, dem Maturitätstyp, dem voraussichtlich gewünschten Studienfach und der Vorbildung der Eltern zu beantworten. Sie wurden je-

doch gebeten, ihr Geschlecht anzugeben und diesen Fragebogen am Testtag mitzubringen und abzugeben.

TEST

Im Vergleich zum Ablauf des Originaltests unter Ernstfallbedingungen (Trost 1977-1994), der in Deutschland an einem Stichtag des Jahres zwischen morgens 08.45 Uhr und nachmittags 16.00 Uhr bearbeitet wird, stand für diesen Probelauf nur ein halber Tag zur Verfügung. Den Gymnasiastinnen und Gymnasiasten wurden deshalb nur 7 von 9 Untertests zur Bearbeitung vorgelegt. Folgender Ablaufplan wurde für den Probelauf festgelegt, wobei die Punkte 1 und 4, respektive 2 und 5 zu je einem Untertest gehören.

	Bearbeitungszeit in Minuten	**Anzahl Aufgaben**
1. Lernphase des Untertests “Figuren Lernen”	4	(20)
2. Lernphase des Untertests “Fakten lernen”	6	(20)
3. Untertest “Textverständnis”	60	24
4. Reproduktionsphase des Untertests “Figuren lernen”	5	20
5. Reproduktionsphase des Untertests “Fakten lernen”	7	20
6. Untertest “Muster zuordnen”	22	24
7. Untertest “Schlauchfiguren”	15	24
8. Untertest “Quantitative und formale Probleme”	60	24
9. Untertest “Konzentriertes und sorgfältiges Arbeiten”	8	1200 Zeichen = 20
Probelauf gesamt	**187**	**156**

Die reine Testzeit betrug demnach 3 Stunden 7 Minuten, und es waren 156 Aufgaben zu bearbeiten. Alle Instruktionstexte wurden aus einem in deutscher und französischer Sprache bereitgestellten Testleiter-Handbuch von den Testleitern zwecks optimaler Durchführungsobjektivität abgelesen.

FRAGEBOGEN 2

Nach der Testdurchführung wurden die Gymnasiastinnen und Gymnasiasten gebeten, zu den unten stehenden Aussagen Stellung zu beziehen. Als Antwortmuster wurde eine vierstufige Skala von “trifft nicht zu”, “trifft wenig zu”, “trifft überwiegend zu” und “trifft genau zu” vorgegeben.

1. Mir ist die Bearbeitung des Tests leicht gefallen.
2. Ich konnte mich die ganze Zeit gut konzentrieren.
3. Ich fühlte mich durch den Test überfordert.

4. Ich fand die Bearbeitung der Aufgaben anstrengend.
5. Ich hatte Angst, am Test teilzunehmen.
6. Ich würde ohne Probleme an einem solchen Test unter Ernstfallbedingungen teilnehmen.

Neben diesen Angaben sollten die Gymnasiastinnen und Gymnasiasten auch noch einschätzen, wieviele Aufgaben in Prozent sie ihrer Erwartung nach richtig gelöst hatten und ob ihre Testleistungen besser/schlechter als die schulischen Leistungen sein würden. Den Getesteten wurde auf diesem Fragebogen 2 auch die Möglichkeit geboten, eigene Bemerkungen zum Test-Probelauf anzubringen. Zugleich konnten sie auch ihre Adresse notieren, falls sie an einer persönlichen schriftlichen Leistungsrückmeldung in Form eines “Testbescheides“ interessiert waren.

KONTROLLGRUPPE

Als Kontrollgruppe wurden die Testpersonen, die 1992 ihren Test für medizinische Studiengänge in Deutschland bearbeitet haben, gewählt. Im Vergleich zu dieser Kontrollgruppe hatte die Stichprobe nicht dieselben Aufgaben zu lösen, sondern vergleichbare Aufgaben mit denselben Aufgabentypen. Die bisher zahlreich vorliegenden Untersuchungen zu den deutschen Testjahrgängen (Trost 1977-1994) zeigen, dass die Resultate der Tests verschiedener Jahrgänge direkt miteinander verglichen werden können. Da die Probelauf-Aufgaben einer Originalversion entsprechen, können auch die Ergebnisse dieses Probelaufs mit den Resultaten des deutschen Originaltests verglichen werden.

DURCHFÜHRUNG

Die Testungen beider Sprachgruppen fanden in separaten Räumen statt. Als erschwerende Bedingung wirkte sich eine klimabedingte hohe Raumtemperatur aus. Da die Räume relativ stark ausgelastet waren, dürfte sich dieser Faktor möglicherweise im Lauf des Tests durch nachlassende Konzentration bemerkbar gemacht haben. Für die Durchführung des Probelaufs bei der französischsprachigen Gruppe wurde ein Testleiter und drei Hilfspersonen zugeteilt. Bei der deutschsprachigen Gruppe wurden ein Testleiter und zwei Hilfspersonen eingesetzt. Insgesamt wurde auf die Einhaltung der Vorgaben des Testleiter-Handbuches geachtet.

Der Probelauf wurde am 28. Juni 1995 von morgens 08.10 Uhr bis mittags 12.00 Uhr durchgeführt. Eine Stichprobe von 54 deutschsprachigen und 126 französischsprachigen Gymnasiastinnen und Gymnasiasten begann den Test. Bei der deutschsprachigen Gruppe haben 12 Personen den Probelauf

vorzeitig abgebrochen. Wegen der erwarteten Motivationsunterschiede waren die Testleiter angehalten, bei Unwillensäusserungen möglichst schnell die Nichtmotivierten den Test nicht weiter bearbeiten zu lassen. Dabei musste in Kauf genommen werden, dass es Gruppentransfer-Effekte gibt. Die Leistungswilligen sollten aber in keinem Falle behindert werden. Die Resultate dieser Abbrecher sind in den Ergebnissen nicht berücksichtigt. Im übrigen dürfte die Quote von Maturandinnen und Maturanden, die sich potentiell überhaupt für ein Medizinstudium - auch aufgrund der Leistungsfähigkeit - bewirbt, nicht über 80% liegen. Bei der Repräsentativerhebung eines Jahrganges dürfte es eine Reihe von Personen geben, die durch den Test objektiv überfordert sind.

Ebenfalls nicht berücksichtigt wurde der letzte Untertest “Konzentriertes und sorgfältiges Arbeiten”. Im Unterschied zu den anderen Tests ist dieser trainierbar und es wird im TEST-INFO empfohlen, den Test vorher mehrmals durchzuführen. Dies ist - möglicherweise vor allem aus Motivationsgründen - in dieser Stichprobe nicht geschehen. Vorhandene Instruktionsunsicherheiten haben dazu geführt, dass die Mehrzahl der Testpersonen die Konzentrationsbogen nicht vorschriftsgemäss ausfüllte (von oben nach unten anstatt von links nach rechts, Kreise oder andere Figuren anstatt Striche, zu schwache Markierungen, Markierung der letzten Lösung etc.). Bei der Nachkontrolle zeigte sich, dass mittels des optischen Lesers nur 50 Prozent der Bogen richtig erfasst worden sind.

Die meisten Untertests enthalten eine Anzahl von Einstreuaufgaben, die für einen zukünftigen Test unter Ernstfallbedingungen vorgetestet werden. In der verwendeten veröffentlichten Originalversion stehen diese Einstreuaufgaben nicht fest. In den Untertests “Muster zuordnen”, “Schlauchfiguren” und “Quantitative und formale Probleme” wurden deshalb zum Zwecke der Vergleichbarkeit mit den deutschen Ergebnissen die vier Items eliminiert, welche die geringste Korrelation zum Untertestergebnis aufwiesen. Entsprechend wurde im Untertest “Textverständnis” ein Aufgabenkomplex mit 6 Items eliminiert.

Ergebnisse

Beschreibung der Stichprobe

Insgesamt haben 42 deutsch- und 125 französischsprachige Gymnasiastinnen und Gymnasiasten (Tabelle 1) den Test mit allen 6 Untertests bearbeitet. Der Anteil der Frauen betrug 45 bzw. 56%. Die entsprechende Anzahl bei der Kontrollgruppe beträgt 23’685 Personen, bei einem Frauenanteil

von 55 Prozent. Das Durchschnittsalter lag bei beiden Schweizergruppen bei 18,4 (Standardabweichung 0,84) und bei der Kontrollgruppe bei 19,8 Jahren (Standardabweichung 2,80). Die Schweizer Stichprobe ist um 1 Jahr jünger als die typische deutsche Teststichprobe.

Gruppe	**Anzahl [n]**	**Anteil Frauen [%]**	**Alter**	
			Mittelwert [m]	**St.abw. [s]**
deutschsprachige Gruppe	42	45	18,4	0,84
französischsprachige Gruppe	126	56	18,4	0,68
Kontrollgruppe BRD	23685	55	19,8	2,80

Tabelle 1: Charakteristika der Testpersonen

Die Testpersonen erhielten für jede korrekt bearbeitete, gewertete Aufgabe einen Punkt. Es ist möglich, im Maximum 118 Punkte (156 - 18 Einstreuaufgaben - 20 vom Konzentrationstest) zu erzielen. Im Test-Probelauf erzielte die beste Testperson 85 Punkte (Abbildung 1). Die niedrigste Punktzahl betrug 30 (bei einem Mittelwert von 56,80 und einer Standardabweichung von 12,33 Punkten). Die entsprechenden Werte für die Kontrollgruppe liegen für nur diese 6 Untertests nicht vor.

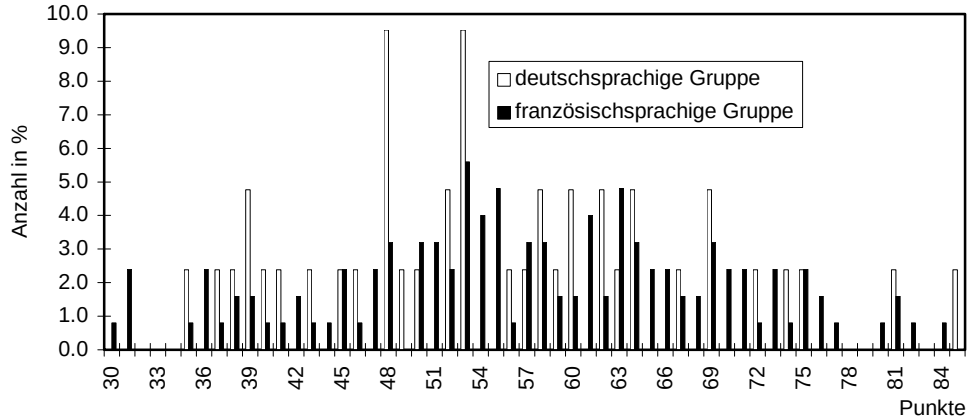


Abbildung 1: Verteilung der Rohwertpunkte in Prozent

Die Verteilung der Rohwertpunkte weist auf die hohe Diskriminationsfähigkeit des Tests hin.

SCHWIERIGKEIT DES TESTS

Der Mittelwert der erreichten Punktzahl liegt in der deutschsprachigen Gruppe bei 55,36 Punkten (Standardabweichung 12,23 Punkte) und in der französischsprachigen Gruppe bei 57,28 Punkten (Standardabweichung 12,38 Punkte). Der Schwierigkeitsindex für den Test-Probelauf beträgt somit 0,47 respektive 0,49. Im Vergleich dazu beträgt er in Deutschland 0,55 (Abbildung 2).

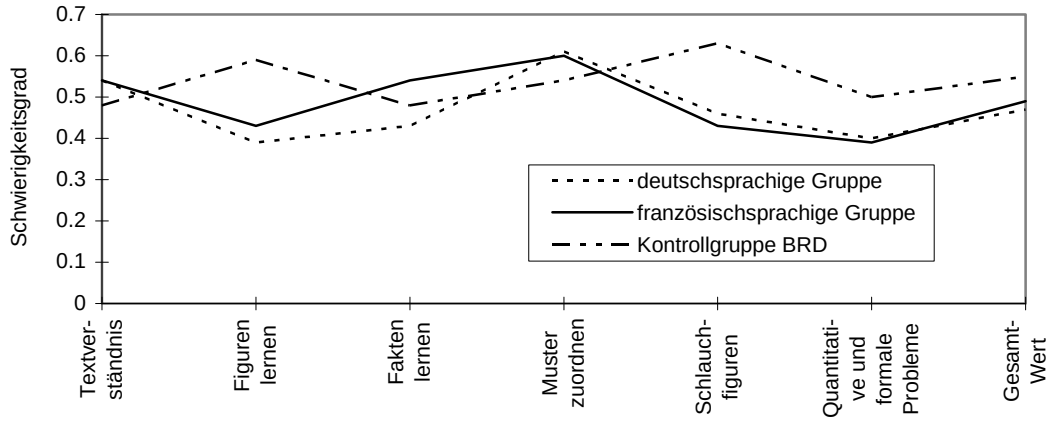


Abbildung 2: Schwierigkeitsgrad des Test-Probelaufs und seiner Untertests

Die mittlere Schwierigkeit der sechs Untertests bewegt sich in der Spanne von 0,39 bis 0,61. Sie liegt somit im angestrebten Bereich von 0,40 bis 0,60, in welchem ausreichend differenziert werden kann. Am schwierigsten fiel der deutschsprachigen Gruppe der Untertest "Figuren lernen" und der französischsprachigen Gruppe der Untertest "Quantitative und formale Probleme"; bei beiden lösten die Testpersonen im Mittel 39 Prozent der gewerteten Aufgaben richtig. Zudem stellte sich in beiden Sprachgruppen der Untertest "Muster zuordnen" als leichtester Test (deutschsprachige Gruppe 0,61; französischsprachige Gruppe 0,60) heraus. In diesem Untertest (Kontrollgruppe 0,54) und im Untertest "Textverständnis" (deutsch- und französischsprachige Gruppe 0,54; Kontrollgruppe 0,48) wurden von den Schweizern mehr Aufgaben richtig gelöst als von der Gruppe aus Deutschland.

MESSGENAUIGKEIT

Der "Cronbach-α-Koeffizient" zeigt, wie gut sich die einzelnen Untertests als Messinstrumente eignen. Abbildung 3 enthält die α-Werte der einzelnen Untertests und den entsprechenden Medianwert für den gesamten Test-

Probelauf. Mit Ausnahme des Untertests “Figuren lernen” (0,52) liegen bei der deutschsprachigen Gruppe die α-Werte um den Median 0,71. Die französischsprachige Gruppe weist im Untertest “Fakten lernen” den niedrigsten α-Wert von 0,46 aus. Ihr Median für den Probelauf beträgt 0,65. Am messgenauesten erwiesen sich für die Schweizer die Untertests “Schlauchfiguren” und “Quantitative und formale Probleme” aus. Dies entspricht den Ergebnissen in Deutschland.

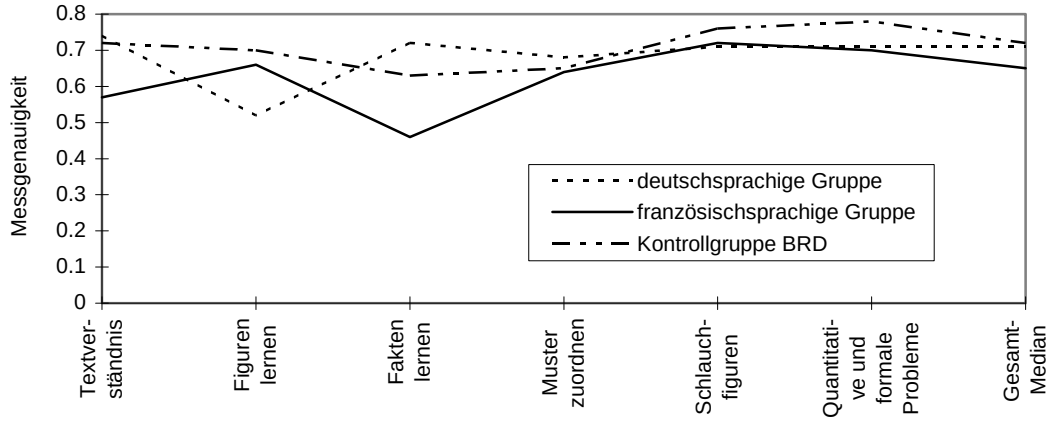


Abbildung 3: Messgenauigkeit des Test-Probelaufs und seiner Untertests

BEZIEHUNGEN ZWISCHEN UNTERTESTS UND ZU SCHULNOTEN

Die Frage, inwieweit sich die Messbereiche der Untertests überschneiden bzw. inwieweit sie jeweils Eigenes messen, ist in Tabelle 2 festgehalten. Die engsten Zusammenhänge unter den Untertests zeigen sich in der Beziehung zum Untertest “Quantitative und formale Probleme”, der auch zum Gesamttest die höchste Korrelation von 0,70 aufweist und zugleich ein auf sehr medizinnahe Inhalte ausgerichteter Test ist.

Die Werte der Korrelationen zwischen den Untertests bzw. des Gesamttests mit den durchschnittlichen Schulnoten bewegen sich in einem Bereich von 0,06 bis 0,23. Das bedeutet, dass die Gesamttestleistung und die Schulleistung im Maximum eine gemeinsame Varianz von 5,3 Prozent aufweisen. Es zeigt sich also, dass der Test etwas anderes misst als die Schule mit ihren Prüfungen.

Untertest	**TV**	**FIL**	**FAL**	**MZ**	**SF**	**QFP**	**GT**
Textverständnis (TV)							
Figuren lernen (FIL)	0.21						

Fakten lernen (FAL)	0.22	0.25					
Muster zuordnen (MZ)	0.14	0.24	0.03				
Schlauchfiguren (SF)	0.26	0.32	0.08	0.30			
Quant. u. form. Probl. (QFP)	0.35	0.15	0.18	0.38	0.40		
Gesamt-Test (GT)	0.59	0.59	0.47	0.58	0.67	0.70	
Durchschnittsnote (NOTE)	0.19	0.17	0.08	0.22	0.07	0.16	0.23

Tabelle 2: Interkorrelationen der Leistungen in den Untertests und deren Korrelationen mit der Schul-Durchschnittsnote

BEZIEHUNG ZWISCHEN GESCHLECHT UND TESTERGEBNIS

Als Anhaltspunkte für substantielle Unterschiede in den Testleistungen werden Effektstärkewerte von $\eta > 0{,}20$ betrachtet. Mittelwertsdifferenzen mit Effektstärkewerten grösser 0,20 sind ebenfalls bei einem t-Test für unabhängige Stichproben signifikant. Im Untertest “Quantitative und formale Probleme” erzielten die Männer in der deutschsprachigen Gruppe im Mittel ein besseres Ergebnis als die Frauen (Tabelle 3). Diese Frauen erreichten dagegen in den Untertests “Figuren lernen”, “Fakten lernen” und “Muster zuordnen” sowie im Gesamttestergebnis ein signifikant besseres Durchschnittsergebnis als ihre männlichen Kollegen. Bei der französischsprachigen Gruppe haben die Frauen in den Untertests “Textverständnis”, “Figuren lernen” und “Fakten lernen” im Mittel ein besseres Ergebnis erzielt als die Männer ihrer Sprachgruppe. In den Untertests “Schlauchfiguren” und “Quantitative und formale Probleme” erreichten dafür jene Männer im Mittel ein signifikant besseres und im Untertest “Muster zuordnen” und im Gesamttestergebnis ein besseres Resultat als ihre Sprachgenossinnen. Verglichen mit der Kontrollgruppe (Tabelle 4) zeigt sich ein heterogenes Bild, und eine Bevorzugung des einen oder des anderen Geschlechts kann nicht festgestellt werden. Mit Ausnahme des Untertests “Textverständnis” weist die französischsprachige Gruppe ähnliche Effekte auf, wie die Kontrollgruppe. Im Untertest “Textverständnis” haben in der Kontrollgruppe die Männer aber in der französischsprachigen Gruppe die Frauen je einen leicht besseren Durchschnittswert erzielt.

	deutschsprachige Gruppe					**französischsprachige Gruppe**				
Untertest / Test	Männer(n=23)		Frauen(n=19)			Männer(n=48)		Frauen(n=60)		
	m	**s**	**m**	**s**	**η**	**m**	**s**	**m**	**s**	**η**

Textverständnis	9.04	3.91	10.37	3.15	0,19	9.60	3.04	9.80	2.83	0,04
Figuren lernen	6.87	2.67	8.84	2.39	0,37	8.04	3.61	8.90	3.24	0,12
Fakten lernen	7.26	2.86	10.26	3.83	0,42	10.52	2.79	11.07	2.99	0,09
Muster zuordnen	11.35	3.52	13.16	3.24	0,26	12.17	3.71	11.73	2.83	0,07
Schlauchfiguren	9.09	3.58	9.11	3.57	0,00	9.92	3.94	7.83	3.43	0,26
Quant.u.form. Probl.	8.26	3.98	7.84	4.36	0,05	9.29	4.26	6.87	3.00	0,31
Gesamt-Test	51.87	11.52	59.58	12.01	0,32	59.54	14.55	56.20	10.84	0,13

Tabelle 3: Beziehung zwischen Geschlecht und Testergebnis in der Stichprobe auf der Test- und der Untertestebene

	Kontrollgruppe BRD				
Untertest	**Männer (n=10'707)**		**Frauen (n=12973)**		
	m	**s**	**m**	**s**	**η**
Textverständnis	9.13	3.68	8.28	3.49	0,12
Figuren lernen	11.41	3.70	11.98	3.67	0,08
Fakten lernen	9.11	3.38	9.91	3.50	0,12
Muster zuordnen	10.74	3.26	10.88	3.22	0,02
Schlauchfiguren	13.43	3.79	11.95	3.92	0,19
Quant.u.form. Probl.	11.32	4.32	8.73	3.83	0,30
Gesamt-Test	65.14	6.98*	61.73	6.67*	0,15*

Tabelle 4: Beziehung zwischen Geschlecht und Testergebnis in der Kontrollgruppe auf der Test- und Untertestebene (* = geschätzt)

Statistische Unterschiede über die jeweils etwa 20 Aufgaben umfassenden Untertests sagen wenig über die Merkmale der einzelnen Items aus. Die einzelnen Itemeffekte können unter Umständen auf der Untertest- oder Testebene verloren gehen (vgl. Klieme 1991). Der Zweck einer detaillierten Betrachtung der Items ist, jene Teile des Tests zu identifizieren, die potentiell zu einer Benachteiligung eines der beiden Geschlechter beitragen könnte. Schwierigkeit, Diskriminationsfähigkeit oder Anfälligkeit für Ratestrategien sind einige Itemkriterien, die zu dieser oder jener Benachteiligung führen können.

Die Untersuchung der Ergebnisse dieser Studie nach auffälligen Items beruht auf einem Verfahren von Mantel & Haenszel (1959). Infolge der klei-

nen Stichprobe besteht jedoch die Gefahr, dass die Unterschiede auch auf Zufall beruhen. Für die Berechnung wurde ein Programm von Rogers & Hambleton (1994) angewendet. Auf Itemebene zeigt sich nun bei der französischsprachigen Gruppe ein Item-Bias zugunsten der Frauen bei der 1. Aufgabe im Untertest "Textverständnis" und bei der 7. Aufgabe im Untertest "Schlauchfiguren" und zugunsten der Männer bei der 14. Aufgabe im Untertest "Schlauchfiguren". Bei der deutschsprachigen Gruppe zeigen sich vier Item-Bias, alle zugunsten der Männer, obwohl diese im Gesamtergebnis im Mittel ein deutlich niedriges Score erzielten als die Frauen ihrer Sprachgruppe. Dabei handelt es sich um das Item 13 vom Untertest "Fakten lernen", das Item 23 vom Untertest "Schlauchfiguren" und die Items 9 und 22 vom Untertest "Quantitative und formale Probleme". Als Beispiel sind in Abbildung 4 alle Schwierigkeitswerte (p-Werte) der Items vom Untertest "Quantitative und formale Probleme", getrennt nach Geschlecht, aufgelistet. Das Liniendiagramm wurde gewählt, weil diese Grafik auf die zunehmenden Itemschwierigkeiten hindeutet und auch eindeutig zeigt, dass man nicht von einer systematischen Benachteiligung des einen oder des anderen Geschlechts sprechen darf.

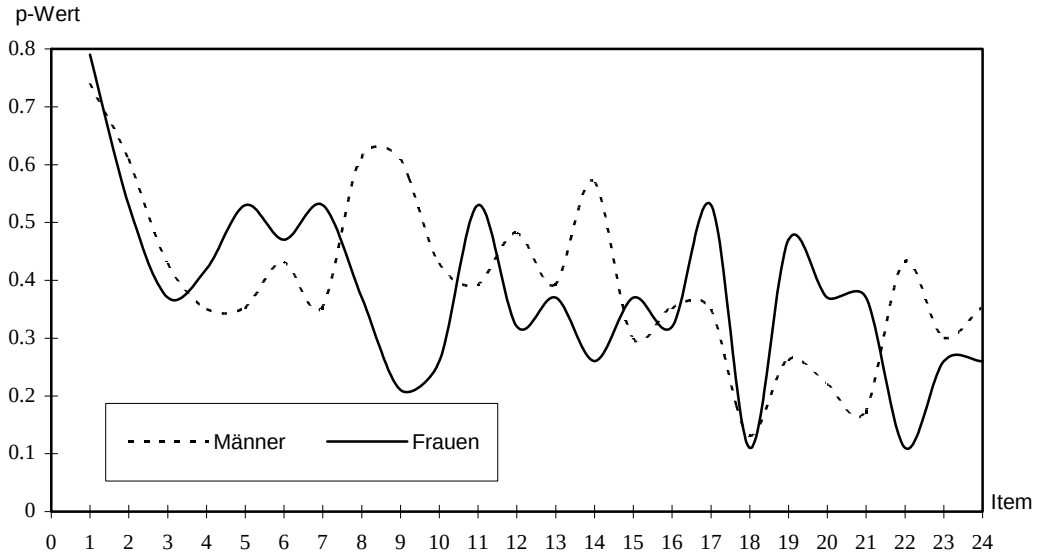


Abbildung 4: Die Beziehung zwischen Geschlecht und Ergebnis im Untertest "Quantitative und formale Probleme" in der deutschsprachigen Stichprobe auf der Itemebene

Von den Items 9 und 22, die einen Bias bezüglich des Geschlechts aufweisen, soll hier beispielsweise anhand des Items 22 gezeigt werden, wie diese detaillierter untersucht wurden. Der Text zu Item 22 mit seinen Trennschärfen je Geschlecht sowie das unterschiedliche Antwortverhalten zwischen den Gymnasiastinnen und Gymnasiasten ist in Tabelle 5 wiedergegeben.

Aufgabe

Eine gewisse Art von Bakterien vermehrt sich in der Weise, dass am ersten Tag aus einem Bakterium zwei, am nächsten aus zwei vier werden usw. Nach 24 Tagen ist das Gefäss, in dem das erste Bakterium angesiedelt wurde, voll.

Hätte man statt einem vier Bakterien angesiedelt, wann wäre das Gefäss voll geworden?

Antwort	**Frauen**		**Männer**	
	n	**%**	**n**	**%**
(A) nach 5 Tagen	3	15,8	1	4,3
(B) nach 6 Tagen	5	26,3	5	21,7
(C) nach 18 Tagen	5	26,3	1	4,3
(D) nach 20 Tagen	2	10,5	3	13,0
(E) nach 22 Tagen	2	10,5	10	43,5
ohne Antwort	2	10,5	3	13,0
Trennschärfe	0,48		0,34	

Tabelle 5: Item 22 des Untertests “Quantitative und formale Probleme”

Weder der Aufgabeninhalt, noch die Lösung (E) zeigen Anhaltspunkte, dass die Frauen der deutschsprachigen Gruppe diese Aufgabe signifikant schlechter lösen konnten als die Männer. Die Trennschärfe bei den Frauen (0,48) zeigt, dass vom Endergebnis her gesehen die Besseren diese Aufgabe richtig gelöst haben, was auch Absicht der Itemkonstruktion ist: nämlich die Testpersonen nach ihrer Testleistung unterscheiden zu können. Es lassen sich auf inhaltlicher Ebene keine schlüssigen Interpretationen finden, die die Entstehung dieses Item-Bias bezüglich des Geschlechts erklären könnten. Zum gleichen Ergebnis kommt man im übrigen bei den anderen Items in beiden Sprachgruppen.

BEZIEHUNG ZWISCHEN SPRACHE UND TESTERGEBNIS

Unterschiede zwischen den Sprachgruppen könnten auf Unterschiede der Übersetzung oder auf Fähigkeitsunterschiede zurückgeführt werden. Der Anteil beider Faktoren kann auf der Basis des hier verwendeten Designs nicht getrennt werden.

Es geht hier nicht um die Feststellung, welche Schweizer Gruppe die besten Ergebnisse erzielt hat, sondern um die Überprüfung der Qualität der Übersetzung in das Französische. Beide Sprachgruppen unterscheiden sich nur im Untertest “Fakten lernen” signifikant voneinander (Tabelle 6). Im Gesamtergebnis hat die französischsprachige Gruppe im Mittel ein leicht besseres Ergebnis erzielt als ihre deutschsprachigen Kolleginnen und Kollegen.

	deutschsprachige Gruppe (n=42)		französischsprachige Gruppe (n=125)		Effekt-stärke
Untertest	**m**	**s**	**m**	**s**	**η**
Textverständnis	9.64	3.61	9.63	2.98	0,00
Figuren lernen	7.76	2.70	8.55	3.47	0,01
Fakten lernen	8.62	3.62	10.75	2.85	0,29
Muster zuordnen	12.17	3.48	11.98	3.19	0,03
Schlauchfiguren	9.10	3.53	8.65	3.77	0,05
Quant.u.form. Probl.	8.07	4.11	7.72	3.67	0,04
Gesamt-Test	55.36	12.23	57.28	12.38	0,07

Tabelle 6: Beziehung zwischen Sprache und Testergebnis

Die Item-Bias-Analyse zwischen den Sprachgruppen zeigt, dass in den sprachunabhängigen Untertests “Muster zuordnen” und “Figuren lernen” drei Items einen Bias aufweisen. Die Items 5 (zugunsten der französischsprachigen Gruppe) und 22 (zugunsten der deutschsprachigen Gruppe) wiesen im sprachabhängigen Untertest “Textverständnis” einen Bias auf. Das gleiche gilt für Item 6 (zugunsten der deutschsprachigen Gruppe) im Untertest “Quantitative und formale Probleme”. Die Abbildung 5 dient der Veranschaulichung der p-Werte, getrennt nach Sprache, des Untertests “Quantitative und formale Probleme”.

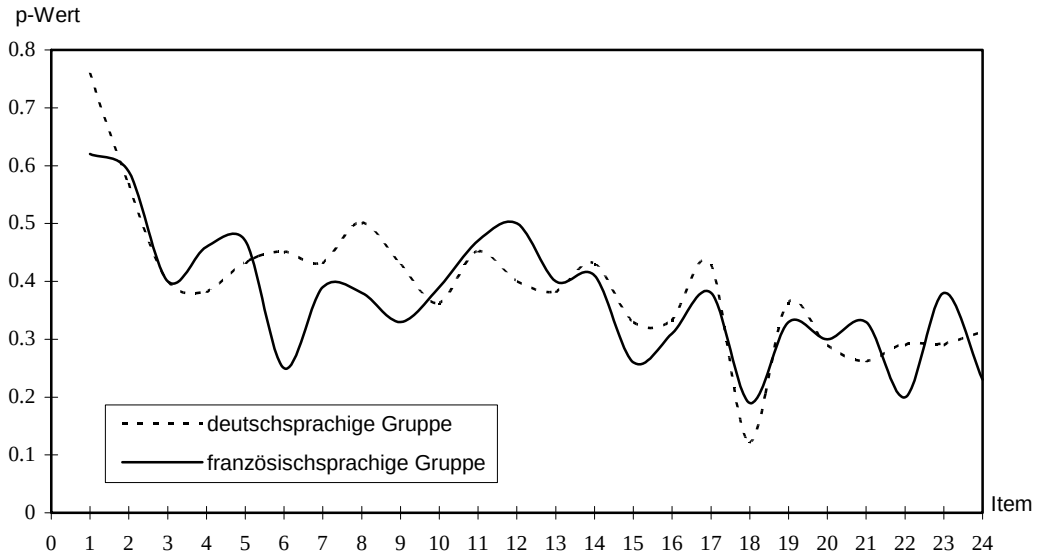


Abbildung 5: Die Beziehung zwischen Sprache und Ergebnis im Untertest "Quantitative und formale Probleme" in der Stichprobe auf der Itemebene

Für beide Sprachgruppen zeigen sich ähnliche Item-Schwierigkeitsgrade. Tabelle 7 enthält Detailangaben zu Item 6 des Untertests "Quantitative und formale Probleme", welches einen Bias infolge der Sprache aufweisen soll. Es macht den Anschein, dass in der französischsprachigen Gruppe eher geraten als gerechnet wurde. Die Lösung (D) ergibt sich nämlich aus dem maximal möglichen "O_2-Schuldvermögen" des Sportlers dividiert durch 5 l O_2 pro Minute (2 ∗ 4 - 3). Dieses Raten führte in dieser Sprachgruppe zu einer geringen Trennschärfe von 0,12. Inhaltlich wurde das Item 6 bei einer weiteren französischen Stichprobe und durch andere als bei der Übersetzung zuständigen zweisprachigen Experten überprüft. Die neuen Erkenntnisse weisen darauf hin, dass es sich um einen statistischen Zufall infolge der kleinen Stichprobe handeln muss. Ebenfalls zum gleichen Ergebnis kam man bei der Überprüfung der anderen Items mit einem Sprach-Bias.

Aufgabe			**Devoir**		
Der menschliche Körper ist in der Lage, für eine kurze Zeit die für die Tätigkeit erforderliche Energie auch dann bereitzustellen, wenn die Sauerstoffaufnahme nicht völlig ausreicht. Es dürfen jedoch insgesamt höchstens 12 l bis 15 l Sauerstoff fehlen ("O_2-Schuld"). Wenn ein Sportler, der ein O_2-Aufnahmevermögen von 3 l pro Minute und ein "O_2-Schuldvermögen" von 14 l hat, eine Tätigkeit erbringt, die 4 l O_2 pro Minute erfordert, dann kann er diese Tätigkeit 14 Minuten durchhalten.			*Le corps humain est capable de mettre à disposition, pour un temps limité, l'énergie nécessaire à une activité même dans les situations où l'oxygénation ne couvre pas entièrement les besoins en O_2. Toutefois, la quantité d'oxygène manquante ne doit pas totaliser plus de 12 à 15 l ("déficit en O_2"). Un athlète ayant une capacité d'oxygénation de 3 l par minute, et supportant un déficit en O_2 de 14 l, est capable de soutenir une activité requérant 4 l d'O_2/minute pendant 14 minutes.*		
Wann wäre die Toleranzgrenze erreicht, wenn der Sauerstoffverbrauch pro Minute auf das doppelte erhöht würde?			*Quand la limite de tolérance serait-elle atteinte si la consommation d'oxygène par minute était doublée?*		
Antwort	**n**	**%**	**réponse**	**n**	**%**
(A) nach 1,7 Minuten	3	7,1	(A) au bout de 1,7 min.	8	6,4
(B) nach 1,8 Minuten	8	19,0	(B) au bout de 1,8 min.	17	13,6
(C) nach 2,5 Minuten	4	9,5	(C) au bout de 2,5 min.	34	27,2
(D) nach 2,8 Minuten	19	45,2	(D) au bout de 2,8 min.	31	24,8
(E) nach 3,5 Minuten	6	14,3	(E) au bout de 3,5 min.	31	24,8

ohne Antwort	2	4,8	sans réponse	4	3,2
Trennschärfe	0,33		sélectivité	0,12	

Tabelle 7: Item 6 des Untertests “Quantitative und formale Probleme” in deutscher und französischer Sprache

TESTERGEBNIS UND WEITERE KRITERIEN

Die weiteren Kriterien wurden infolge der kleinen Stichprobe nur noch auf der Testebene untersucht. Bei der Stichprobe handelt es sich um die homogene Altersgruppe eines Jahrganges (Tabelle 8, siehe auch Tabelle 1). Für die Überprüfung der Beziehung zwischen Alter und Testergebnis wurde die Stichprobe in die „Unter-Neunzehn-Jährigen“ und in die „Über-Achtzehn-Jährigen“ aufgeteilt. Bei der deutschsprachigen Gruppe erzielten im Mittel die „Unter-Neunzehn-Jährigen“ weniger Punkte als ihre älteren Sprachgenossen, bei der französischsprachigen Gruppe ist es umgekehrt.

	deutschsprachige Gruppe			französischsprachige Gruppe		
Alter	**n**	**m**	**s**	**n**	**m**	**s**
< 19	25	52.84	13.34	69	59.32	12.21
> 18	13	58.85	10.93	37	55.35	13.09
η	0,23			0,15		

Tabelle 8: Beziehung zwischen Alter und Testergebnis

Die Prüflinge konnten gleichzeitig mehrere Schul-Interessensgebiete auf dem Fragebogen 1 angeben. Trotzdem ergab dies zu kleine Gruppen für die einzelnen Interessensschwerpunkte. Deshalb wurden diese Interessensschwerpunkte (Tabelle 9) zu den Interessensgebieten Sprache (altsprachlich und neusprachlich), Naturwissenschaft (Biologie, Chemie und Physik), Mathematik und Übrige (Geographie, Wirtschaft, Sport, Kunst, etc.) zusammengefasst. Es zeigte sich nun, dass in der deutschsprachigen Gruppe im Mittel diejenige Interessensgruppe das beste Ergebnis erzielte, welche in der Schule vorwiegend an den naturwissenschaftlichen Fächern interessiert ist. Ihnen folgen die Interessensgruppen Mathematik, Übrige und Sprache. Bei der französischsprachigen Gruppe liegt die Interessensgruppe Mathematik leicht vor der Interessensgruppe Naturwissenschaft, gefolgt von Übrige und Sprache.

Interessensschwerpunkt	**deutschsprachige Gruppe**			**französischsprachige Gruppe**		
	n	**m**	**s**	**n**	**m**	**s**
Sprache	16	51.56	10.98	60	57.32	11.50
Naturwissenschaft	13	57,15	15,93	52	61,50	12,19
Mathematik	18	55,89	12,69	35	62,00	12,00
Übrige	23	54.09	13.19	78	57.44	12.96

Tabelle 9: Beziehung zwischen Schul-Interessensschwerpunkt und Testergebnis

Die 3 Prüflinge der deutschsprachigen Gruppe vom Maturitätstyp B haben im Mittel in ihrer Sprachgruppe das beste Testergebnis erzielt (Tabelle 10). Der Scheffé-Test für die Durchführung von A-posteriori-Vergleichen ergibt in dieser Sprachgruppe keine signifikanten Mittelwertsunterschiede.

Dagegen weisen in der französischsprachigen Gruppe die Prüflinge vom Maturitätstyp D im Mittel ein signifikant niedrigeres Ergebnis auf als die Personen der Maturitätstypen A, B und C. In dieser Sprachgruppe erzielte die Maturitätstyp-C-Gruppe im Mittel die meisten Punkte.

Maturitätstyp	deutschsprachige Gruppe			französischsprachige Gruppe		
	n	**m**	**s**	**n**	**m**	**s**
A				13	59.23	11.93
B	3	57.33	4.51	33	57.61	11.10
C	23	55.96	14.15	35	63.26	10.21
D	12	52.25	11.43	17	46.71	10.56
η	0,15			0,41		

Tabelle 10: Beziehung zwischen Maturitätstyp und Testergebnis

	französischsprachige Gruppe			
Maturitätstyp	**A**	**B**	**C**	**D**
A				*
B				*
C				*
D				

Scheffé-Test für die französischsprachige Gruppe; * signifikanter Unterschied ($p < 0,05$) zwischen den Mittelwerten der Maturitätstypen A, B und C zu dem Mittelwert des Maturitätstypus D

Bei der Angabe der zukünftigen Studienfachwahl (Tabelle 11), welche die Prüflinge des Test-Probelaufs in Zukunft gedenken zu wählen, lässt sich feststellen, dass in beiden Sprachgruppen die noch nicht Entschlossenen im Mittel das beste Ergebnis erzielt haben. Es folgen bei der deutschsprachigen Gruppe diejenigen, welche voraussichtlich nicht studieren wollen, und bei der französischsprachigen Gruppe diejenigen, die wissen, welches Fach sie studieren wollen, das aber nicht Medizin ist. Der Scheffé-Test ergibt keine signifikanten Mittelwertsunterschiede zwischen den einzelnen Studienfachwahl-Gruppen.

voraussichtliche Studienfachwahl	**deutschsprachige Gruppe**			**französischsprachige Gruppe**		
	n	**m**	**s**	**n**	**m**	**s**
Medizin	1	50.00	0.00	8	55.75	10.01
Nicht Medizin	15	51.40	14.78	30	58.40	13.79
noch offen	17	57.88	10.28	54	58.59	11.63
kein Studium	5	56.20	14.97	5	50.00	13.98
Effektstärke η	0,25			0,15		

Tabelle 11: Beziehung zwischen voraussichtlicher Studienfachwahl und Testergebnis

Die Prüflinge des Test-Probelaufs konnten auf dem Fragebogen 1 eine Angabe über die Vorbildung ihrer Eltern machen. Infolge der kleinen Stichprobe wurden die verschiedenen Ausbildungsabschlüsse zusammengefasst zu Abschluss 1 (nicht bekannt oder keinen Abschluss, Realschule, Sekundarschule und Berufslehre) und Abschluss 2 (Matura, höhere Fachschule, Lehrer- bzw. Lehrerinnenseminar, Universität oder Hochschule und Promotion) (Tabelle 12). Es zeigt sich, dass Jugendliche der französischsprachigen Gruppe, deren Mütter einen Abschluss 2 erreicht haben, im Probelauf im Mittel ein besseres Ergebnis erzielten als die anderen Jugendlichen ihrer Sprachgruppe. Bei der deutschsprachigen Gruppe ist weder ein Einfluss der Vorbildung des Vaters noch der Mutter zu erkennen; das gilt auch für die Vorbildung des Vaters in der französischsprachigen Gruppe.

Vorbildung		**deutschsprachige Gruppe**			**französischsprachige Gruppe**		
Eltern		**n**	**m**	**s**	**n**	**m**	**s**
Abschluss 1	Vater	15	54,67	12,29	45	57,69	13,62
Abschluss 2	Vater	21	55,81	13,70	44	57,57	12,05
Effektstärke η		0,04			0,02		
Abschluss 1	Mutter	24	55,00	11,78	54	55,37	13,17
Abschluss 2	Mutter	11	55,82	16,35	34	60,88	11,64
Effektstärke η		0,03			0,18		

Tabelle 12: Beziehung zwischen Vorbildung der Eltern und Testergebnis

Es mögen unterschiedliche Gründe dazu geführt haben, dass die Prüflinge des Test-Probelaufs sich durch den Test überfordert fühlten, oder dass es für sie eher schwierig war, sich über die ganze Zeit gut konzentrieren zu können. In der französischsprachigen Gruppe haben diejenigen, die sich weniger überfordert und konzentrierter erlebten, im Mittel ein signifikant besseres Testresultat erzielt als die jeweilige Komplementärgruppe. Bei den anderen Meinungen, wurden keine Mittelwertsunterschiede in den Testergebnissen festgestellt (Tabelle 13).

Meinungen zum Test	**deutschsprachige Gruppe**			**französischsprachige Gruppe**		
	n	**m**	**s**	**n**	**m**	**s**
Bearbeitung: eher leicht	10	56.40	13.07	12	60.00	14.10
Bearbeitung: eher schwierig	32	55.03	12.16	105	57.30	12.30
Effektstärke η	0,05			0,07		
Konzentration: eher gut	17	55.59	12.17	31	62.10	13.23
Konzentr.: eher schlecht	25	55.20	12.52	88	56.24	11.70
Effektstärke η	0,02			0,21		
Überforderung: eher ja	24	53.17	10.66	72	54.18	10.91
Überforderung: eher nein	16	56.63	13.76	45	62.80	13.30
Effektstärke η	0,15			0,33		
Anstrengung: eher ja	36	55.53	11.92	98	57.64	12.28
Anstrengung: eher nein	5	50.40	13.13	18	57.61	14.55
Effektstärke η	0,14			0,03		
Angst: eher ja	1	62.00	0.00	7	60.86	14.25
Angst: eher nein	40	55.33	12.46	109	57.30	12.19
Effektstärke η	0,08			0,07		

Tabelle 13: Beziehung zwischen Meinungen der Testpersonen und Testergebnis

Bei beiden Schweizer Gruppen nutzten jeweils etwas über die Hälfte die Möglichkeit, eine Rückmeldung anzubringen (Tabelle 14). In den Kommentaren wurden vor allem der ungünstige Testtermin, die zu schlecht

belüfteten Räume und die vermisste Pause thematisiert. Ebenfalls die fehlende Motivation wurde angesprochen, und es wurde deshalb bezweifelt, ob die Ergebnisse für einen Probelauf aussagekräftig seien. Es zeigte sich nun, dass in beiden Sprachgruppen die Testpersonen, die einen Kommentar abgegeben haben, im Mittel weniger Punkte erzielten als die übrigen Testpersonen.

	deutschsprachige Gruppe			**französischsprachige Gruppe**		
Kommentare	**n**	**m**	**s**	**n**	**m**	**s**
ja	23	54.09	12.92	68	55.22	11.83
nein	19	56.89	11.49	57	59.74	12.67
Effektstärke η	0,12			0,18		

Tabelle 14: Beziehung zwischen Kommentar ja oder nein und Testergebnis

Nur 32 Prozent der deutsch- und 36 Prozent der französischsprachigen Gruppe waren der Ansicht, dass sie über 50 Prozent der Aufgaben richtig gelöst haben (Tabelle 15). Im Mittel wurden diese Prognosen bei beiden Sprachgruppen auch erreicht (deutschsprachige Gruppe: 51.2%; französischsprachige Gruppe: 53.3%). Dazu erzielten bei beiden Sprachgruppen die “Über-49-Prozentigen” im Mittel ein signifikantes besseres Ergebnis als die “Unter-50-Prozentigen”.

Prozent richtiger Lösungen	**deutschsprachige Gruppe**			**französischsprachige Gruppe**		
	n	**m**	**s**	**n**	**m**	**s**
< 50 %	25	53.32	11.55	54	54.11	11.01
> 49 %	12	60.42	12.69	31	62.90	13.11
Effektstärke η	0,26			0,28		

Tabelle 15: Beziehung zwischen Einschätzung der prozentual richtigen Lösungen und Testergebnis

Nur einige der Testpersonen (deutschsprachige Gruppe: 5; französischsprachige Gruppe: 6) gaben an, dass ihre Testleistungen besser sein werden als ihre Schulleistungen (Tabelle 16). Der entsprechende Teil der deutschsprachigen Gruppe erzielte im Mittel auch ein besseres Ergebnis (58.00) als das Pendant ihrer Sprachgruppe (55.17). Das umgekehrte Ver-

hältnis in den Notendurchschnitten (Tabelle 17) scheinen die Angaben über das Test-Schulleistungsverhältnis in dieser Sprachgruppe zu bestätigen. Da nur eine der sechs Personen in der französischsprachigen Gruppe ihre Durchschnittsnote angegeben hat, lassen sich hier keine weiteren Ergebnisse ableiten.

	deutschsprachige Gruppe Testleistungen			**französischsprachige Gruppe Testleistungen**		
Testleistung	**n**	**m**	**s**	**n**	**m**	**s**
besser als Schulleistung	5	58.00	12.81	6	48.50	9.33
schlechter als Schulleistung	35	55.17	12.58	102	57.45	12.01
Effektstärke η	0,08			0,16		

Tabelle 16: Beziehung zwischen Einschätzung der Test- im Vergleich zur Schulleistung und Testergebnis

	deutschspr. Gruppe Schulleistungen			**französischspr. Gruppe Schulleistungen**		
Testleistung	**n**	**m**	**s**	**n**	**m**	**s**
besser als Schulleistung	3	4.23	0.25	1	4.50	0.00
schlechter als Schulleistung	29	4.88	0.34	76	4.72	0.29
Effektstärke η	0,51			0,12		

Tabelle 17: Beziehung zwischen Einschätzung der Test- im Vergleich zur Schulleistung und Testergebnis

Die Testpersonen, die sich eine Leistungsrückmeldung gewünscht haben, erzielten ein signifikant besseres Ergebnis als die Testpersonen, die auf eine Rückmeldung verzichteten (Tabelle 18). Demzufolge darf davon ausgegangen werden, dass die an einer Rückmeldung interessierte Gruppe versucht hatte, ein möglichst gutes Ergebnis zu erzielen.

	deutschspr. Gruppe			**französischspr. Gruppe**		
Leistungsrückmeldung	**n**	**m**	**s**	**n**	**m**	**s**
ja	15	60.07	11.61	52	62.73	10.88

nein	27	52.74	11.97	73	53.40	11.96
Effektstärke η	0,29			0,37		

Tabelle 18: Beziehung zwischen gewünschter Leistungsrückmeldung und Testergebnis

Diskussion

Die Durchführung eines Probelaufs mit dem Eignungstest, der als Selektionskriterium für den geplanten Numerus Clausus im Medizinstudium vorgesehen ist, konnte wie vorgesehen erfolgen. Die Probleme der Organisation einer Testdurchführung können mit den vorhandenen Strukturen gelöst werden. Einige Erfahrungen während der Durchführung werden Eingang in das Testleiter-Handbuch finden, wo alle Abläufe genau beschrieben werden. Ein Training der Testleiter scheint sehr notwendig. Die Raumgrösse sollte 100 Personen nicht überschreiten. Da die Testung im Juli stattfinden würde, ist auf klimatisch geeignete Räume ganz besonders zu achten.

DER TEST ERFÜLLT DIE GÜTEKRITERIEN

Trotz fehlender Bewerbungsmotivation und einer Durchführung unter nicht optimalen Raumbedingungen wurden ähnliche Gütekriterien wie in Deutschland erzielt. Bezüglich der Messgenauigkeit schnitten die Schweizer Stichproben in 2 von 6 Untertests im Mittel leicht besser ab als die deutsche Kontrollgruppe. Die Niveau-Unterschiede lassen sich nach der Analyse der Testdaten wohl vor allem auf die fehlende Bewerbungsmotivation, die dem Probelauf zugrunde lag, zurückführen. Die Zuverlässigkeit und die Testfairness entsprechen in beiden Sprachgruppen insgesamt aber den Anforderungen an ein psychodiagnostisches Verfahren. Vor allem die Zuverlässigkeit des Testgesamtwertes, der für die Zulassung verwendet würde, entspricht in beiden Sprachgruppen im Niveau dem der deutschen Kontrollgruppe.

DER TEST ERLAUBT EINE OPTIMALE DIFFERENZIERUNG

Die Rohwertverteilungen, die Schwierigkeitsgrade und die Trennschärfen erlauben die Differenzierung der Kandidatinnen und Kandidaten nach ihrer Testleistung. Es wurde im Vergleich zur deutschen Kontrollgruppe (55%) im Mittel ein Schwierigkeitsgrad von 48% erreicht. Entsprechend der Si-

tuation in der Schweiz wäre eine Differenzierung derart nötig, dass zwischen 75 und 85 Prozent der Testbesten zum Studium zugelassen werden könnten. Der Test müsste also in diesem Bereich ausreichend gut differenzieren. Es haben in diesem Bereich jeweils maximal 2,3 Prozent der Personen den gleichen Punktwert erreicht. Ein Grenz-Punktwert, den nur die entsprechend der Kapazität festgelegte Quote überschreitet, liesse sich also hinreichend genau finden.

DER TEST MISST NICHT DAS GLEICHE WIE DIE SCHULNOTEN

Auf der Basis der Schulnoten konnte auch berechnet werden, inwieweit sich Test- und Schulleistungen entsprechen. Die relativ niedrige Korrelation von 0,23 des Testwertes mit einer (erfragten) Gesamtnote weist daraufhin, dass der Test nicht genau das gleiche misst, was in den Schulnoten zum Ausdruck kommt. Entsprechende Korrelationen in Deutschland fallen etwas höher aus (0,39 - vgl. Trost et al. 1994). Bei der Erklärung dieses Unterschiedes zwischen Schulnoten und Testergebnis weisen Trost et al. auf korrelationsmindernde Einflüsse unterschiedlicher Beurteilungsmassstäbe hin, die für Schulnoten gelten können. Vor allem das Problem der in der Schweiz von Kanton zu Kanton und von Schule zu Schule uneinheitlichen Bewertungsstrenge bei der Vergabe der Maturitätsnoten könnte sich hier weiter korrelationsmindernd auswirken. Demgegenüber sind die Bewertungsmassstäbe des Tests einheitlich und es wäre ein Ausgleich dieser Unterschiede möglich. Dabei ist in Rechnung zu stellen, dass die prognostische Validität des Tests für den Studienerfolg den Schulnoten keinesfalls unterlegen ist.

Die Überprüfung der Beziehungen zwischen den Untertests belegt im übrigen, dass jeder Untertest für sich genommen eigene Kriterien testet und damit keiner der Untertests überflüssig - weil redundant - ist.

Bei der Analyse der Beziehungen zwischen Test und weiteren Kriterien erwies sich auch die Beziehung zwischen Maturitätstyp und Testergebnis als bedeutsam. In der französischsprachigen Gruppe erzielten die Personen vom Maturitätstyp D ein signifikant niedrigeres Durchschnittsergebnis als die Personen, die ihre Matura in den Maturitätstypen A bis C abschliessen. Dies würde vergleichbaren Ergebnissen im Medizinstudium entsprechen: Beispielsweise an der Medizinischen Fakultät der Universität Bern bestehen zwischen 44% und 50% der Absolventinnen und Absolventen von Maturitätstyp A bis C das 1. Propaedeutikum im ersten Anlauf nicht, bei Typ D sind es dagegen 77% (Hofer 1992).

DER TEST BENACHTEILIGT FRAUEN NICHT

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. In einigen Darstellungen kam der Test in den Verdacht, mehr oder weniger frauendiskriminierend zu sein. Bei der Bewertung des Tests wurde dort allerdings immer über die deutschen Ergebnisse gesprochen, die hier allerdings nicht bestätigt werden konnten.

Die Resultate des Probelaufs zeigen keine systematischen Unterschiede für die beiden Geschlechter. Weder auf der Test-, der Untertest- noch der Itemebene konnte nachgewiesen werden, dass man von möglichen Benachteiligungen der Frauen in der Schweiz ausgehen muss. Nur 4 von 136 Items in der deutschsprachigen und 3 von 136 Items in der französischsprachigen Gruppe wiesen überhaupt einen Bias bezüglich des Geschlechts auf. Auf der Untertestebene erzielten die Frauen der deutschsprachigen Gruppe im Mittel in 5 von 6 Untertests ein besseres Resultat als die Männer in ihrer Sprachgruppe (3 davon sind statistisch signifikant). In der französischsprachigen Gruppe verlief der Untertestvergleich ausgeglichen. Während die Frauen in den Untertests “Textverständnis”, “Figuren lernen” und “Fakten lernen” ein besseres Durchschnittsergebnis erreichten, war es in den Untertests “Muster zuordnen”, “Schlauchfiguren” und “Quantitative und formale Probleme” gerade umgekehrt. Auf der Testebene erreichten bei der deutschsprachigen Gruppe die Frauen und bei der französischsprachigen Gruppe die Männer im Mittel die besseren Ergebnisse als ihre Geschlechtsgenossen. Verglichen mit der deutschen Kontrollgruppe kann festgestellt werden, dass sich die deutschsprachige Gruppe geradezu umgekehrt verhält.

Die mögliche Feststellung von Unterschieden in einer Testung würde im übrigen nicht automatisch bedeuten, dass dieser Unterschied auch so in das Zulassungskriterium übernommen wird. Wenn der politische Entscheid dazu gefasst wird, kann ein statistisches Korrekturverfahren angewendet werden, welches den Mittelwert-Unterschied zwischen den Geschlechtern ausgleicht. Die Zulassungsquoten können genau den Bewerbungsquoten für beide Geschlechter angeglichen werden. Die Anwendung dieses Verfahrens erfordert allerdings neben der politischen Entscheidung auch den tatsächlichen Nachweis, dass die im Eignungstest genannten Unterschiede in der Schweiz auftreten würden.

CHANCENGLEICHHEIT GILT AUCH FÜR SPRACHGRUPPEN

Der Originaltest wurde in Deutschland konstruiert. Bei der Übernahme eines solchen Tests in eine andere Kultur und Sprache können Unterschiede auftreten, welche die Chancengleichheit beeinflussen. Deshalb muss schon bei der die Hin- und Rückübersetzung der Items durch zwei unabhängige Expertenteams darauf geachtet werden, dass keine Verluste und Verzerrungen der Inhalte entstehen. Nach der Durchführung des adaptierten Tests kann mit der Überprüfung der Gütekennwerte festgestellt werden, ob Mängel durch die Adaption entstanden sind.

Es wiesen nur 6 von 136 Items einen Bias bezüglich der Sprache auf, wovon 3 aus eher sprachunabhängigen Untertests stammen. Bei den übrigen Items wurden die Trennschärfen, die Schwierigkeitswerte, die Aufgabeninhalte und das Antwortverhalten in den einzelnen Sprachgruppen untersucht. Diese Unterschiede sind allerdings eher unsystematisch. In den Untertests “Textverständnis”, “Muster zuordnen”, “Schlauchfiguren” und “Quantitative und formale Probleme” hat die deutschsprachige Gruppe ein leicht besseres Durchschnittsergebnis erzielt als die französischsprachige Gruppe. Letztere hat jedoch in den Gedächtnistests “Figuren lernen” und “Fakten lernen” im Mittel soviel Punkte mehr erreicht, dass sie auch auf der Testebene ein Plus gegenüber der anderen Sprachgruppe aufweisen kann. Die sorgfältige und aufwendige Übersetzung des Tests in die französische Sprache hat also dazu geführt, dass man von einer gelungenen Adaption des Tests in eine andere Kultur sprechen kann.

FAZIT

Alle **Voraussetzungen** für die Anwendung eines Eignungstests für das Medizinstudium in der Schweiz sind **gegeben**, wenn die politische Entscheidung für seinen Einsatz getroffen wird.

Die schweizerdeutsche und die französischsprachige Adaptation des Tests erreichen bezüglich **Zuverlässigkeit und Testfairness** eine der deutschen Originalfassung gut vergleichbare Testgüte.

Auf der Basis der Testgesamtwerte kann die Zulassung zum Medizinstudium mit **hinreichender Differenzierung** erfolgen. Die Schwierigkeit des Tests für beide Schweizer Sprachgruppen unterscheidet sich nur unwesentlich von derjenigen in der deutschen Gruppe.

Es konnte **nicht bestätigt** werden, dass in der Schweiz von **Unterschieden** in den Testwerten bezüglich der **Geschlechter** auszugehen ist. Damit wäre eine Chancengleichheit für Frauen und Männer von vornherein gegeben.

Die **Chancengleichheit** ist auch für den Vergleich der beiden **Sprachgruppen** vorhanden. Es konnte damit auch unter Beweis gestellt werden, dass das gewählte Adaptationsverfahren angemessen ist.

Ein Vorteil des Tests ist es, dass er **nicht Schulwissen** prüft, sondern die Eignung zum Medizinstudium erfasst.

Der Probelauf hat bestätigt, dass die **organisatorisch-technischen Voraussetzungen** für seine Anwendung in der Schweiz **erfüllt** werden können.

Literatur

Hänsgen, Klaus-Dieter, Hofer, Rainer & Ruefli, Daniel (1995). Der Eignungstest für das Medizinstudium in der Schweiz. Grundlagen, Anwendung und Probleme. Schweizerische Ärztezeitung, 76(37), 1476-1496.

Hofer, Rainer (1992). Die Beziehung zwischen Maturitätstyp und Erfolg im 1. Propaedeutikum an der medizinischen Fakultät der Universität Bern. Unveröffentlichte Studie. Universität Bern: Institut für Aus-, Weiter- und Fortbildung.

Institut für Test- und Begabungsforschung (Hrsg.) (1990). Test für medizinische Studiengänge. Aktualisierte Originalversion 2. Göttingen: Hogrefe.

Klieme, Eckhard (1991). Problemstellung: Fairness von TMS-Aufgaben. In G. Trost (Hrsg.) Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 15. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Rogers, H. Jane & Hambleton, Ronald K. (1994). MH: A fortran 77 program to compute the Mantel-Haenszel statistic for detecting differential item functioning. Educational and Psychological Measurement, 54(1), 101-104.

Trost, Günter (Hrsg.) (1977-1994). Test für medizinische Studiengänge (TMS): Studien zur Evaluation. 1. - 18. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung.

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

Zentrum für Testentwicklung und Diagnostik (Hrsg.) (1995). Test-Info. Eignungstest für das Medizinstudium in der Schweiz. Information für die Anmeldung 1995. Universität Fribourg.