Reliabilität des Hamburger Auswahlverfahrens für Medizinische Studiengänge, Naturwissenschaftsteil (HAM-Nat)

Johanna Hissbach[1]
Dietrich Klusmann[2]
Wolfgang Hampe[1]

1 Universitätsklinikum Hamburg-Eppendorf, Institut für Biochemie und molekulare Zellbiologie, Hamburg, Deutschland

2 Universitätsklinikum Hamburg-Eppendorf, Institut und Poliklinik für Medizinische Psychologie, Hamburg, Deutschland

Zusammenfassung

Ziele: Die Universität Hamburg hat im Jahr 2005 begonnen, einen Naturwissenschaftstest zur Auswahl von Studienbewerbern zu entwickeln (Hamburger Auswahlverfahren für Medizinische Studiengänge, Naturwissenschaftsteil, HAM-Nat). Diese Studie ist ein weiterer Schritt, den HAM-Nat zu etablieren. Wir untersuchen

1. die Paralleltest- und Retest-Reliabilität,
2. die Auswirkungen eines Chemiekurses auf die Testergebnisse, sowie
3. die Übereinstimmung der Testergebnisse des HAM-Nat mit denen des Testmoduls „Naturwissenschaftliches Denken“, das inhaltlich und strukturell dem Modul „Medizinisch-naturwissenschaftliches Grundverständnis“ des Tests für Medizinische Studiengänge (TMS) entspricht.

Methoden: 316 Studienanfänger nahmen an der Studie in der Orientierungseinheit im Jahr 2007 teil. Sie bearbeiteten verschiedene Versionen des HAM-Nat, die jeweils aus alten Fragen (HN2006) und neuen Fragen (HN2007) bestanden. Nach vier Wochen bekam die eine Hälfte der Studienanfänger erneut den HAM-Nat, allerdings nur die 2007er Version; die andere Hälfte bekam das Modul „Naturwissenschaftliches Denken“. Innerhalb dieser 4 Wochen konnten die Studienanfänger an einem fünftägigen Chemiekurs teilnehmen.
Ergebnisse: Die Paralleltest-Reliabilitäten für die vier Testversionen lagen zwischen r_{tt}=.53 und r_{tt}=.67. Die Retest-Reliabilitäten der beiden 2007er Testhälften lagen bei r_{tt}=.54 und r_{tt}=.61. Die HAM-Nat Versionen HN2006 und HN2007 korrelierten mit dem Modul „Naturwissenschaftliches Denken“ zu r=.34 und r=.21. Studierende, die zwischen Test und Testwiederholung einen Chemiekurs absolviert hatten, verbesserten dadurch nicht ihre Testleistungen.
Schlussfolgerungen: Die Ergebnisse lassen erwarten, dass weitere Testversionen zu naturwissenschaftlichem Wissen ebenfalls keine hohe interne Konsistenz, Paralleltest-Reliabilität oder Retest-Reliabilität ergeben. Daher ist für den Aufbau einer Sammlung von Items, die austauschbar für die Erzeugung von Parallelversionen benutzt werden können, große Sorgfalt erforderlich. Das Testmodul „Naturwissenschaftliches Denken“ misst im Wesentlichen etwas anderes als der HAM-Nat. Die Tatsache, dass die Teilnahme an einem Chemiekurs keinen Effekt auf die Leistungen im Chemie-Teil des HAM-Nat hatte, ist vermutlich auf fehlende Abstimmung der Inhalte des Kurses mit dem Test zurückzuführen und auf die geringe Motivation der Studienteilnehmer, besonders zum zweiten Testzeitpunkt.

Schlüsselwörter: Studienbewerberauswahl Medizin, Externe Validität, Reliabilität, Studieneingangstest

Einleitung

Auf der Suche nach einem geeigneten Auswahlverfahren für das Medizinstudium entwickelt die Universität Hamburg seit dem Jahr 2005 einen Naturwissenschaftstest (HAM-Nat) zur Studienbewerberauswahl. Hintergrund hierfür ist die Änderung der Gesetzeslage, die es den Universitäten erlaubt, 60% ihrer Studierenden selbst auszuwählen [1]. In Hamburg dürfen laut Gesetz zur

Studienbewerberauswahl unter anderem schriftliche Auswahltests eingesetzt werden [2].

Vor 2008 wurden die Studienbewerber in Hamburg allein nach ihrer Abiturdurchschnittsnote ausgewählt. Dies ist ein einfaches Verfahren, und die Abiturnote hat sich als brauchbarer Prädiktor für Studienleistungen bewährt. Für die Kohorten von Medizinstudierenden von 1986/1987 fanden Trost et al. [3] eine Korrelation von r=0.48 für die Abiturnote mit dem Ergebnis des schriftlichen Teils des Physikums. Mit dem mündlichen Teil betrug die Korrelation r=.34 [3]. In ihrer Metaanalyse berichten Trapmann et al. [4] eine korrigierte prädiktive Stärke von r=.58 für Studiennoten im vorklinischen Studienabschnitt. Auch in ausländischen Studien [5] und in nichtmedizinischen Fächern [4] besitzen Schulabschlussdurchschnittsnoten eine hohe prognostische Validität. In einer prospektiven englischen Studie zeigte sich eine gewisse Vorhersagekraft von Schulabschlussnoten in Bezug auf die Berufsausübung von Ärzten [6].

Dennoch wird die Auswahl nach Abiturnote immer wieder kritisiert. Trapmann et al. [4] fassen zusammen:

1. geringe Vergleichbarkeit der Abiturnoten zwischen den verschiedenen Schulen und Bundesländern,
2. unzureichende Reliabilität und Validität von Schulnoten,
3. unterschiedliche Bewertungsmaßstäbe für verschiedene Klassen und von verschiedenen Lehrern.

Die Vorhersagekraft der Abiturdurchschnittsnote für den Studienerfolg sinkt in späteren Abschnitten des Studiums. Weil sich sehr viele Abiturienten zum Medizinstudium bewerben, liegt der zur Zulassung erforderliche Notendurchschnitt auf einem hohen Niveau. Bewerber, die Hamburg mit erster Ortspräferenz wählten, mussten in den Jahren 2005 – 2007 einen Notendurchschnitt von mindestens 1,6-1,7 aufweisen, um zugelassen zu werden. Gerade weil die Abiturnoten in den verschiedenen Bundesländern auf verschiedenen Schulformen, Fächerkombinationen und Bewertungsmaßstäben basieren, wirft die Abiturnote als alleiniges Kriterium Fragen der Fairness auf [7]. Die Hinzunahme weiterer Auswahlkriterien kann die Nachteile der Abiturnote teilweise ausgleichen.

Einige deutsche Fakultäten setzen zur Ergänzung der Abiturnote den Test für Medizinische Studiengänge (TMS) ein, der zwischen 1986 und 1996 für alle Studienbewerber der Medizin verbindlich war. Dieser Test enthält zwar naturwissenschaftliche Fragen, zielt aber auf ein anderes Konstrukt: spezifische Studierfähigkeit [8]. Die Korrelationen von Abiturnote und TMS-Ergebnis zwischen r=.37 bis r=.48 deutet darauf hin, dass Schul- und Testleistung hinreichend unterschiedliche Leistungsaspekte erfassen [3]. Die Vorhersagekraft des TMS beruht im Wesentlichen auf vier medizinnahen Aufgabengruppen (medizinisch-naturwissenschaftliches Grundverständnis, Lösung quantitativer und formaler Probleme, Textverständnis, Verständnis von Diagrammen und Tabellen) [3].

Kenntnistests, die den Wissensstand in studienfachrelevanten Bereichen prüfen, werden bereits in vielen Ländern verwendet [9]. Unter anderem in Belgien [10] und Österreich [11] werden medizinspezifische Kenntnistests für die Studierendenauswahl eingesetzt. Reibnegger et al. [12] zeigten, wie nach Einführung eines Auswahlverfahrens gegenüber dem offenen Zugang die Anzahl der Studierenden, die in der Regelstudienzeit das Grundstudium absolvierten, von 23% auf 84% der Studierenden anstieg (Mittelwerte der 3 Jahre vor und nach der Einführung). Die Abbruchrate unter den Studienanfängern im ersten Studienjahr sank von 10% bei offenem Zugang auf 1% nach Einführung des Auswahlverfahrens. Der überwiegende Teil des Tests bestand aus naturwissenschaftlichen Fragen, ähnlich den Fragen des HAM-Nat.

In England wird seit 2003 an einigen Universitäten der Biomedical Admissions Test (BMAT) zur Bewerberauswahl eingesetzt. Der Wissensteil des Tests („scientific knowledge and application“) erwies sich als brauchbarer Prädiktor der Examensleistungen im ersten und zweiten Studienjahr [13]. Die Prädiktion durch den 2. Teil des BMAT, in dem ebenfalls mit Multiple-Choice Fragen Problemlösung, Textverständnis und die Interpretation von Daten und Grafiken überprüft wird („aptitude and skill“), ist deutlich schlechter [14].

In Deutschland gibt es neben dem HAM-Nat gegenwärtig kein Auswahlverfahren mit spezifisch naturwissenschaftlichem Inhalt für medizinische Studiengänge. Mit dem Kenntnistest für Naturwissenschaften HAM-Nat führen wir auch in Hamburg ein zweites Qualifikationskriterium neben der Abiturnote ein, das einheitlich für alle Bewerber gilt und dessen Testeigenschaften fortlaufend untersucht werden können. Der HAM-Nat soll naturwissenschaftliche Kenntnisse prüfen, die für den Erfolg im ersten Studienabschnitt wichtig sind. Damit sollen Bewerber ausgewählt werden, die eine gute Chance haben, erfolgreich zu studieren. Zugleich soll der Test die Möglichkeit geben, eine schwächere Abiturnote auszugleichen. Seit 2008 finden Studienbewerber auf der Homepage des Universitätsklinikums Eppendorf eine Internetseite mit Themenkatalog und Selbsttest (http:// www.uke.uni-hamburg.de/ studienbewerber). Die Internetseite hat nicht nur das Ziel, über das Studium zu informieren und einen realistischen Test für naturwissenschaftliche Kenntnisse anzubieten, sondern sie soll auch die Studienbewerber dazu anhalten, ihre Motivation zum Studium und ihre Fähigkeit, es erfolgreich zu absolvieren, selbst zu prüfen. Gewünscht ist eine Selbstselektion, die der Selektion durch die Universität vorangeht. Vorbereitung auf den HAM-Nat ist zugleich auch Vorbereitung auf das Studium, denn die naturwissenschaftlichen Fragen des HAM-Nat prüfen genau das Wissen, auf dem die naturwissenschaftlichen Studienfächer aufbauen.

In einer Pilotstudie im Jahr 2006 wurden die ersten HAM-Nat Items zunächst Oberstufenschülern mehrerer Gymnasien vorgelegt. Daraus entstand die erste Testversion für die Studienanfänger der Kohorte 2006 [15]. Für eine weitere Voruntersuchung des Tests wurden für die Kohorte 2007 neue Items erzeugt. Damit stellt sich die Frage, ob die neue 2007er Testversion zu der 2006er Version parallel ist.

Die vorliegende Untersuchung soll diese Frage beantworten und darüber hinaus die Retest-Reliabilität prüfen. Weiterhin untersuchen wir den Effekt eines Lernprogramms (fünftägiger Trainingskurs) in Chemie auf die Testleistung und die Übereinstimmung des HAM-Nat mit dem Testmodul „Naturwissenschaftliches Denken“, das inhaltlich und strukturell dem TMS Subtest „medizinisch-naturwissenschaftliches Grundverständnis“ entspricht.

Methoden

Testentwicklung HAM-Nat

Einen Überblick über die Vorarbeiten zur Entwicklung der 2006er Version des HAM-Nat liefern Hampe et al. [15]. Nachdem 8 Items, die in der Vortestung an Gymnasiasten wenig trennscharf waren, entfernt worden waren, bestand der HN2006 aus 52 Items. Zu diesen Items erzeugte eine Arbeitsgruppe 60 inhaltlich und formal ähnliche Testfragen für einen Paralleltest – den HN2007. Diese 2007er Version des HAM-Nat besteht aus 60 Multiple-Choice Fragen aus medizinrelevanten Themengebieten der Fächer Mathematik, Chemie, Physik und Biologie auf dem Niveau der gymnasialen Oberstufe. Die Arbeitsgruppe bestand aus Gymnasiallehrern sowie von Dozenten der klinischen und theoretischen Fächer der medizinischen Fakultät.

Beispiel für eine HAM-Nat Frage:

Bei der Oxidation eines Aldehyds entsteht ...
A) ein Ester.
B) ein Keton.
C) eine Carbonsäure.
D) ein Alkohol.
E) ein Alken.

Eine der fünf Antwortalternativen ist jeweils richtig. Die Teilnehmer hatten pro Frage 1,5 Minuten Zeit zur Bearbeitung. Der aktuelle Themenkatalog sowie Fragen aus den Jahren 2006 und 2007 sind als Selbsttest auf der Internetseite des Universitätsklinikums Eppendorf (UKE) zu finden (http:// www.uke.uni-hamburg.de/ studienbewerber).

Testmodul „Naturwissenschaftliches Denken“

Die Aufgabengruppe „Naturwissenschaftliches Denken“ ähnelt inhaltlich und strukturell dem Modul „medizinisch-naturwissenschaftliches Grundverständnis“ des Tests für Medizinische Studiengänge (TMS). Beide Tests wurden von der ITB-Consulting GmbH entwickelt. Das Testmodul beinhaltet 24 Multiple-Choice Aufgaben, die mit der Schilderung eines naturwissenschaftlichen Sachverhalts beginnen. Es werden verschiedene Behauptungen aufgestellt und der Testteilnehmer muss entscheiden, ob diese Behauptungen den vorangegangenen Beschreibungen nach richtig sind. Es gibt jeweils 5 Antwortalternativen, von denen eine richtig ist. Die Bearbeitungszeit ist auf 55 Minuten begrenzt. Die Aufgaben setzen kein spezifisch naturwissenschaftliches Wissen voraus, sondern zielen auf die Durchdringung eines Sachverhalts und die Fähigkeit zu schlussfolgerndem Denken ab. Das Recht, das Testmodul durchzuführen, wurde von der ITB-Consulting GmbH erworben.

Chemiekurs

Der fünftägige Chemiekurs wird regelhaft für die Studienanfänger der Medizin nach der Orientierungseinheit, aber vor Beginn des ersten Semesters am Fachbereich Chemie der Universität Hamburg angeboten. Ziel des Kurses ist, das unterschiedliche Vorwissen der Studierenden anzugleichen. Die Teilnahme ist freiwillig, die Durchführung tutorengestützt. Es werden mehrere parallele Kurse in Gruppengrößen von 30-40 Studienanfängern angeboten, in denen Themen der gymnasialen Oberstufe, wie z.B. der Materiebegriff, der Begriff der chemischen Reaktion und organische Verbindungen und deren Aufbau zunächst vom Tutor vorgestellt und anschließend in Übungsaufgaben bearbeitet werden. Die Inhalte des Kurses ähneln denen des HAM-Nat-Themenkataloges. Die Tutoren kannten diesen jedoch nicht und bereiteten die Teilnehmer nicht gezielt auf den HAM-Nat vor.

Studiendesign

1. Testzeitpunkt: Paralleltests

Der 2006er Test bestand aus zwei Testhälften A und B mit jeweils 26 Items. Da der Test zuvor im Internet veröffentlicht worden war, konnten diese Items den Probanden bekannt sein, sofern sie die Seiten besucht hatten. Der 2007er Test bestand aus den Testhälften C und D mit jeweils 30 Items, die neu entwickelt worden waren. Die Studienteilnehmer bearbeiteten jeweils zwei Testhälften (AC, AD, BC oder BD), nämlich 26 alte Fragen aus dem HN2006 und 30 neue Fragen aus dem HN2007 (siehe Abbildung 1). Vor dem 2. Testzeitpunkt bestand die Möglichkeit, an dem freiwilligen fünftägigen Chemiekurs teilzunehmen. Die Anzahl der Tage, an denen die Studienanfänger am Kurs teilnahmen, wurde erfragt.

2.Testzeitpunkt: Retest und Testmodul „Naturwissenschaftliches Denken“ nach 4 Wochen

Die Studienteilnehmer wurden randomisiert in zwei Gruppen aufgeteilt. Vier Wochen nach der ersten Testung bearbeiteten 96 Testteilnehmer den kompletten HN2007, also die Testhälften C und D. Das bedeutet, dass sie eine Testhälfte schon kannten, während die andere für sie neu war. Eine Woche später bearbeitete die andere Hälfte der Studienteilnehmer (N=91) das Modul „Naturwissenschaftliches Denken“. Die Durchführung des Tests im Anschluss an eine Pflichtlehrveranstaltung wurde

durch Mitarbeiter unserer Arbeitsgruppe organisiert und von Dozenten der Medizinischen Fakultät beaufsichtigt. Der Test „Naturwissenschaftliches Denken“ wurde eigens für die Studie durchgeführt, unabhängig von den offiziellen, bundesweit angebotenen TMS-Terminen für die Studierendenauswahl.

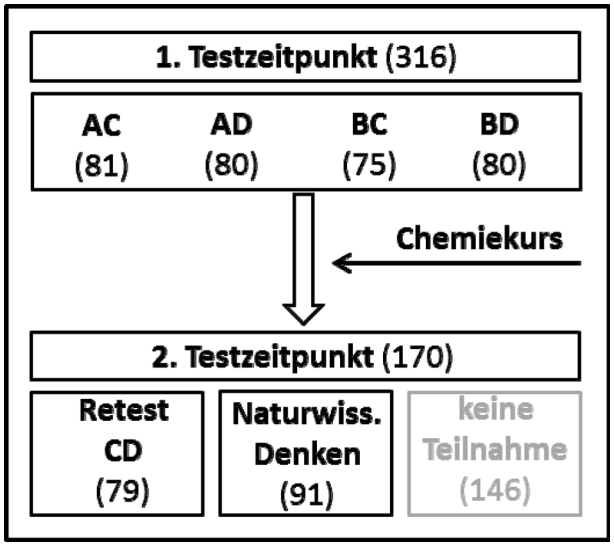

Abbildung 1: Studiendesign und Anzahl der Studienteilnehmer

Stichprobe

Allen Studienanfängern der Medizin, die in der ersten Semesterwoche 2007 an der Orientierungseinheit der Universität Hamburg teilnahmen, wurde die Teilnahme am Test angeboten. Die Teilnahme war freiwillig, alle Probanden willigten schriftlich in die Verwendung ihrer Daten ein. Die Stichprobe setzt sich folgendermaßen zusammen (siehe Abbildung 1): Für die Auswertung der Paralleltestreliabilität (1. Testzeitpunkt) liegen die HAM-Nat Daten von 316 Personen vor (77% der gesamten Kohorte). Die Stichprobe bestand aus einem Drittel Männer und zwei Drittel Frauen. Dies entspricht der Geschlechterverteilung der gesamten Kohorte. Die mittlere Abiturdurchschnittsnote war 1,8. Beim Wiederholungstermin (2. Testzeitpunkt) nahmen 170 Personen (54 % der Ausgangstichprobe) am Test teil, denen ein Ergebnis aus der ersten Testung zugeordnet werden konnte. Der zweite Testtermin war an ein Seminar mit verpflichtender Teilnahme gekoppelt. Anders als in der Orientierungseinheit waren viele Anwesende nicht bereit, ein zweites Mal am Test teilzunehmen. Der Vergleich von Testwiederholern und Teilnahmeverweigerern ergab keine signifikanten Unterschiede in der Abiturnote und der Geschlechterverteilung. Zum zweiten Testzeitpunkt bearbeiteten 91 Teilnehmer das Modul „Naturwissenschaftliches Denken“ und 79 erneut den HAM-Nat. Der Effekt des Chemiekurses kann an 52 Studierenden evaluiert werden, die den HAM-Nat vor und nach dem Kurs bearbeiteten und Angaben zur Teilnahme am Kurs machten. Nicht alle Teilnehmer absolvierten den kompletten Kurs, 15 Personen gaben an, mit 3 oder weniger Tagen und 37 mit mehr als 3 Tagen am Kurs teilgenommen zu haben.

Statistische Auswertung

Die Parallelität von Testformen drückt sich in gleichen wahren Werten und gleichen Fehlervarianzen aus. Anhaltspunkte für die Parallelität verschiedener Testversionen sind gleiche Mittelwerte und Streuungen, sowie eine hohe Korrelation zwischen den Testformen. Der Retest-Reliabilität liegt die Annahme zugrunde, dass sich zwischen zwei Messzeitpunkten die wahren Werte der Testteilnehmer nicht verändern und dass die Einflüsse von Messfehlern konstant sind. Sie bezeichnet den Grad der Übereinstimmung der Ergebnisse eines bestimmten Tests für dieselben Probanden bei wiederholten Messungen. Als Maß für die Paralleltest-Reliabilität und für die Übereinstimmung der HAM-Nat Ergebnisse mit dem Modul „Naturwissenschaftliches Denken“ und der Abiturnote wählten wir die Pearson Korrelation, für die Retest-Reliablitität des HN2007 Spearmans Rangkorrelation.
Cronbach’s α ist der Erwartungswert für die Korrelation zweier Itemsets mit dem Umfang k, die nach Zufall aus dem Universum aller möglichen Items (für das gegebene Konstrukt) ausgewählt wurden. Wenn die Tests HN2006 und HN2007 parallel sind, dann müssen die Korrelationen zwischen den Teilskalen aus HN2006 und den Teilskalen aus HN2007 ebenso hoch sein wie ihre internen Konsistenzen. Wenn die Korrelationen unterschiedlich sind, dann greifen entweder beide Tests nicht auf das gleiche Universum möglicher Items zu oder sie haben zwar ein Universum möglicher Items gemeinsam, sind aber keine Zufallsauswahlen daraus.
Für die Analyse der einzelnen Testhälften wurden die Summenscores des HN2006 und des HN2007 als Messwiederholung (Innersubjektfaktor) und die Gruppenzugehörigkeit (AC, AD, BC, BD) als Zwischensubjektfaktor im Allgemeinen Linearen Modell betrachtet. Ein signifikanter Messwiederholungseffekt bedeutet, dass die Testversionen unterschiedlich schwierig sind, Interaktionseffekte mit Gruppenzugehörigkeit geben Auskunft über die Unterschiedlichkeit der beiden Hälften innerhalb eines Tests.
Um den Effekt des Chemiekurses auf die Testleistung in Items aus dem Fach Chemie zu untersuchen, wurde die Teilnahme am Kurs als dichotome Variable (0 bis 3 Tage vs. mehr als 3 Tage) als Zwischensubjektfaktor in ein neues Modell mit aufgenommen, in dem die Fragen nach Fachgebiet (Chemiefragen vs. andere Fragen) und Testzeitpunkt (erste Testung vs. Retest) getrennt als Innersubjektfaktoren behandelt wurden. Für die Analysen wurde PASW 18 für Windows [16] verwendet.

Ergebnisse

Interne Konsistenz und Paralleltest-Reliabilität

Die Inter-Item-Korrelationen lagen für alle Skalen zwischen r=-.22 und r=.53 (Mittelwert: .06), die internen

Konsistenzen der Testhälften lagen zwischen α=.56 und α=.69 (siehe Tabelle 1) und die Paralleltest-Korrelationen zwischen r=.53 und r=.67 (siehe Tabelle 2).

Tabelle 1: Skalenstatistiken der vier verschiedenen Testversionen

Testversion	HN2006		HN2007	
	A	B	C	D
gültige N	162	155	157	160
Anzahl der Items	26	26	30	30
Cronbach's Alpha (KR-20)	.687	.630	.628	.557
Mittlere Itemschwierigkeit	.440	.451	.405	.351
Mittlere Itemvarianz	.206	.204	.204	.199
Mittlere Inter-Item-Korrelation	.076	.060	.052	.037
Mittlere Item-Score-Korrelation	.229	.195	.181	.147
Skalenmittelwert	11.45	11.73	12.14	10.53
Skalenvarianz	15.72	13.42	15.58	12.94
Korrelation mit Abiturnote	-.278	-.126	-.221	-.335

Tabelle 2: Korrelation der verschiedenen Testhälften

		HN 2007	
		Testhälfte C	Testhälfte D
HN 2006	Testhälfte A	.674	.660
	Testhälfte B	.531	.532

Retest-Reliabilität

Die Retest-Reliabilität wurde nur für die Testversion HN2007 berechnet. Für Testhälfte C betrug die Rangkorrelation r_{tt}=.52 (n=46), für Testhälfte D r_{tt}=.61 (n=34) (siehe Abbildung 2). Die entsprechenden Pearson Korrelationen waren r_{tt}=.54 und r_{tt}=.56. Einige Testteilnehmer schnitten im Retest schlechter ab als bei ihrer ersten Testung (siehe Abbildung 2). Wenn die 9 Teilnehmer, die in der zweiten Testung in einer der beiden Testhälften weniger als 6 Punkte erreicht hatten, aus der Rechnung ausgeschlossen wurden, erhöhte sich die Korrelation nicht (Testhälfte C r_{tt}=.45, n=39; Testhälfte D r_{tt}=.61, n=32), obwohl Abbildung 2einen solchen Effekt suggerieren mag.

Unterschiede zwischen den Testversionen 2006 und 2007

Eine detailliertere Weise, die Unterschiede zwischen den Tests zu betrachten, bietet das Allgemeine Lineare Modell (ALM). Im ALM mit den Faktoren Testversion (HN2006 vs. HN2007) als Messwiederholungsfaktoren und Testhälfte (A oder B bzw. C oder D) als Zwischensubjektfaktor zeigte sich, dass von den 2007er Fragen signifikant weniger gelöst wurden als von den alten 2006er Fragen (38.5% vs. 45.2%, $F_{1,312}$=101.5; p<.001). Während alle Testteilnehmer in den beiden 2006er Testhälften etwa gleiche Ergebnisse erzielen ($F_{1,312}$=2.3; p=.128), ist die Testhälfte D ist mit 35.1% gelösten Fragen etwas schwieriger als Testhälfte C mit 40.6% gelösten Fragen ($F_{1,312}$=11.4; p=.001). Wird der Zwischensubjektfaktor Geschlecht in das Modell aufgenommen, zeigt sich kein signifikanter Einfluss des Geschlechts auf die Leistungen in den verschiedenen Testversionen (F=.468, p=.495), obwohl die Männer in ihrer Gesamtleistung im Test besser abschneiden als die Frauen (44% vs. 40% richtige Antworten, T=-2.64; p=.009).

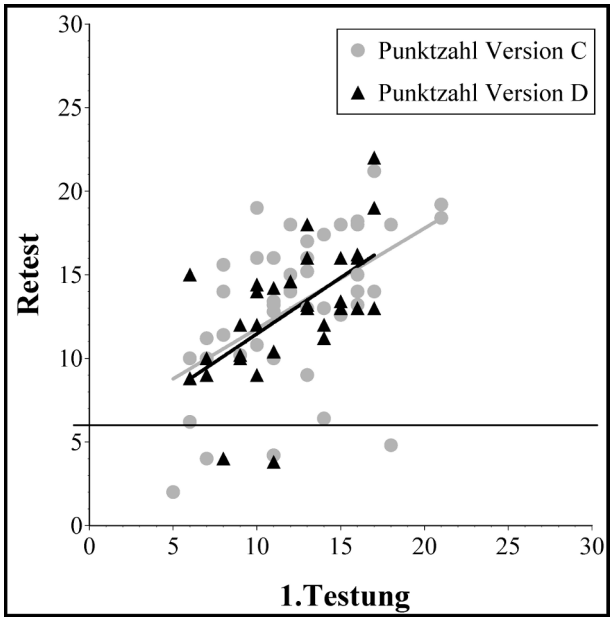

Abbildung 2: Test/ Retest-Korrelationen für Testhälften C und D des HN2007 und die jeweiligen Regressionsgeraden. Unter der waagerechten Gerade liegen alle Testteilnehmer, die in der zweiten Testung Ergebnisse unterhalb der Ratewahrscheinlichkeit erzielten.

Effekt des Chemiekurses

Um den Effekt eines Chemiekurses zu untersuchen, wurde für die Chemiefragen und die übrigen Fragen (Biologie, Physik, Mathematik) getrennt untersucht, inwieweit sich die Ergebnisse in den beiden Testungen vor und nach dem Kurs unterscheiden. Die Chemiefragen wurden weniger häufig richtig beantwortet als die restlichen Fragen zu den Themengebieten Biologie, Physik und Mathematik (35.8% vs. 43.4% richtig Antworten, $F_{78,1}$=25.6, p<.001). Es gab weder eine Verbesserung noch eine Verschlechterung der HN2007-Ergebnisse nach dem Chemiekurs ($F_{1,78}$=0.26; p=.611), auch nicht für die Chemieitems (Interaktionseffekt: $F_{1,78}$=0.26; p=.610).
Um den Einfluss der Dauer der Teilnahme am Kurs zu untersuchen, wurde die Variable „Intensität der Teilnahme am Kurs" dichotomisiert in 0-3 Tage vs. 4-5 Tage. Da nicht alle Testteilnehmer angaben, ob sie am Kurs teilgenommen hatten, reduziert sich die Stichprobe auf n=52. Auch die Intensität der Teilnahme am Kurs hatte keinen signifikanten Effekt auf die gesamte Leistung im HN2007 ($F_{1,50}$=2.4; p=.124) oder die Leistung in den Chemieitems ($F_{1,50}$=0.1; p=.759). Wurde das Geschlecht als weiterer

Faktor in das Modell aufgenommen, ergaben sich keine signifikanten Interaktionseffekte (alle p>.289).

Bekanntheit der Fragen

Bei der zweiten Testung war die eine Hälfte der Fragen für die Testteilnehmer bekannt, die andere Hälfte war neu. Die bekannten Fragen wurden in der zweiten Testung (41.5 %) nicht signifikant häufiger richtig beantwortet als in der ersten (40.1 %; $F_{1,50}$=0.4; p=.543). Auch der Ausschluss von Testpersonen, die in der zweiten Testung sehr schlechte Leistungen zeigten, änderte nichts an diesen Ergebnissen.

Korrelation Abiturnote und HAM-Nat

Die Korrelation der Abiturdurchschnittsnote mit den unterschiedlichen HAM-Nat Versionen lag zwischen r=-.34 und r=-.13 mit einem Mittelwert von r=-.24 (siehe Tabelle 1). Die Korrelation des Moduls „Naturwissenschaftliches Denken“ mit der Abiturnote betrug r=-.11 (n=90).

Korrelation mit dem Modul „Naturwissenschaftliches Denken“

Das Testmodul „Naturwissenschaftliches Denken“ korrelierte mit der Testhälfte A des HN2006 zu r=.34 und mit der Testhälfte B ebenfalls zu r=.34. Die Korrelationen mit den Testhälften des HN2007 lagen bei r=.19 für Version C und r=.23 für die Version D (siehe Abbildung 3). Für die zusammengefassten Testhälften betrugen die Korrelationen mit dem Modul „Naturwissenschaftliches Denken“ r=.34 (HN2006, A+B) und r=.21 (HN2007, C+D). Die Korrelationen unterschieden sich nicht signifikant (p=.350, Testung mit Fisher‘s z [17]).

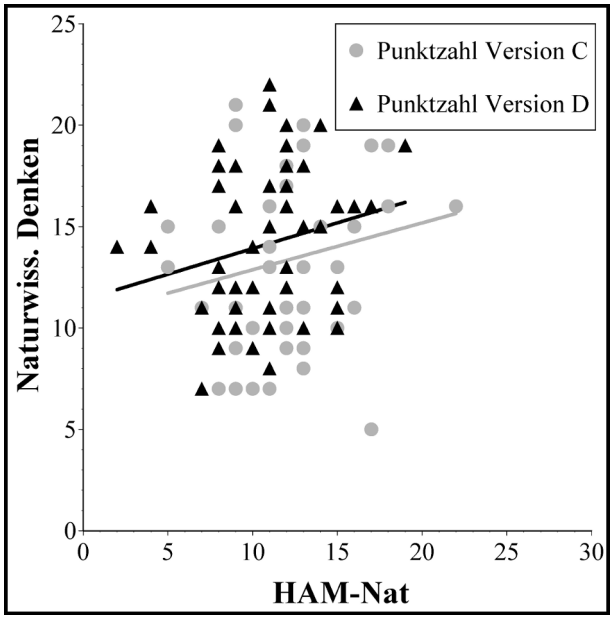

Abbildung 3: Korrelationen des Moduls Naturwissenschaftliches Wissen aus dem TMS mit den jeweiligen Testhälften des HAM-Nat und die jeweiligen Regressionsgeraden

Diskussion

Die Ergebnisse zur Frage der Parallelität beider Testversionen lassen sich folgendermaßen zusammenfassen:

1. Im neuen Test (HN2007) wurden signifikant weniger Items gelöst als im alten (HN2006) und die Wiederholung des gleichen Tests nach vier Wochen führte nicht zu besseren Testleistungen.
2. Der alte und der neue Test unterschieden sich nicht signifikant bezüglich ihrer interner Konsistenzen und ihren Korrelationen mit einem dritten Test, dem Modul „Naturwissenschaftliches Denken“.
3. Die internen Konsistenzen (Cronbach’s α) der jeweils aus den Testversionen gebildeten Testhälften unterschieden sich nicht signifikant von den Korrelationen der Testhälften (Paralleltest-Reliabilität).

Warum ist HN2006 leichter als HN2007? Möglich wäre, dass einige Studierende die Internetdarstellung des HN2006 kannten und dadurch einen Vorteil hatten. Doch dieser Effekt kann nicht sehr stark sein, denn die Studierenden waren bereits zugelassen und wenn sie sich mit dem HN2006 im Internet beschäftigt hatten, dann nicht, um sich auf eine ernsthafte Prüfung vorzubereiten. Wir wissen nicht, wie viele Probanden die Seite besucht haben. Zum Vergleich: Die Wiederholung der Testung mit Halbformen des HN2007 ergab nicht die geringste Verbesserung trotz des kurzen Zeitintervalls von vier Wochen zwischen Test und Retest. Warum sollte dann der vermutlich seltene und kursorische Besuch einer Internetseite einen Effekt haben? Wahrscheinlicher ist, dass 2007 die Erzeuger der Testfragen tatsächlich schwierigere Items produziert haben.

Einerseits stellen die unterschiedlichen Schwierigkeiten kein Problem für den HAM-Nat dar, weil der Zweck dieses Tests darin besteht, Bewerber in eine Rangreihe zu bringen, um die Zulassung zum Studium in Kombination mit anderen Faktoren (Abiturnote, weitere Tests) zu regeln. Solange zwei Tests dieselbe Rangreihe produzieren, sind sie auch austauschbar. Andererseits sollte ein Test, der als Auswahlkriterium herangezogen wird, ein greifbares Profil besitzen und seine Beschaffenheit nicht unkontrolliert von Jahr zu Jahr ändern.

Ein Maß dieser Reproduzierbarkeit ist die Rangkorrelation. Sie beträgt für Testhälften C des HN2007 r=.52 und für die Testhälfte D r=.61. Das sind keine hohen Werte, wenn man bedenkt, dass vier Wochen nach der ersten Testung die gleichen Items vorgelegt wurden. Der Grund für die geringe Reproduzierbarkeit der Rangreihe ist vermutlich ein Störfaktor, der für die gesamte Untersuchung gilt: Da es sich um keine echte Bewerbungssituation handelte, reflektieren die Testwerte nicht nur Wissensunterschiede sondern auch Motivationsunterschiede. Dies betrifft besonders den 2. Testzeitpunkt, zu welchem die Studienteilnehmer durch die Anforderungen des Studienbeginns stark gefordert waren. Hier hat nur noch knapp mehr als die Hälfte der Ausgangsstichprobe teilgenommen. Die niedrige Retest-Korrelation sollte daher als eine Unterschätzung betrachtet werden.

Die besonders schlechten Leistungen in Chemie könnten sich dadurch erklären lassen, dass der Chemieunterricht an den meisten Schulen erst später eingeführt wird als die anderen Naturwissenschaften und zudem häufiger in der Oberstufe abgewählt wird. Diese Schüler haben daher ein sehr viel geringeres Chemiewissen im Vergleich z.B. zum Biologiewissen von Schülern, die dieses Fach in der Oberstufe abgewählt, zuvor jedoch bereits viele Jahre ein Grundwissen erworben hatten. Daher ist es sinnvoll, einen Trainingskurs anzubieten, um die Wissenslücken im Fach Chemie zu schließen. Warum aber spiegelte sich die Teilnahme am Chemiekurs nicht in besseren Leitungen im Chemieteil des HAM-Nat? Dieser Teil der Studie ist besonders auf die Motivation beim Wiederholungstest angewiesen, die, wie oben beschrieben, wahrscheinlich nicht sehr hoch war. Möglicherweise erfassten aber auch die HAM-Nat-Items teilweise ein Wissen, das im Kurs nicht behandelt wurde. Auch dieser Befund lenkt die Aufmerksamkeit auf den Vorgang der Itemerzeugung. Neue Items sollten mit dem typischen Lehrmaterial korrespondieren, das Bewerber für ihre Vorbereitung benutzen. Nur so kann Vorbereitung die Chance auf Zulassung tatsächlich verbessern – eine der gewünschten Wirkungen des HAM-Nat. Zur Verbesserung des HAM-Nat im Jahr 2008 wurde daher ein Themenkatalog veröffentlicht, um den Studienbewerbern die Vorbereitung auf den Test zu erleichtern. Alle Fragen des 2008er HAM-Nat können eindeutig einem oder mehreren Themengebieten des Katalogs zugeordnet werden.

Die gegenüber dem HN2006 nicht signifikante, aber leicht geringere interne Konsistenz des HN2007 könnte dadurch erklärbar sein, dass bei dieser Version keine Vorselektion von Items nach Trennschärfe stattfand wie bei der Version HN2006. Um das zu prüfen, haben wir für beide Versionen Items ausgeschlossen, die Trennschärfen <.10 aufwiesen und die interne Konsistenzen neu berechnet. In den 2006er Testhälften lagen nur 5 Items unter .10, während es für die beiden Testhälften des HN2007 insgesamt 15 waren. Wurden diese Items eliminiert, lagen die internen Konsistenzen für alle Testhälften zwischen .60 und .70. Damit sind die internen Konsistenzen nur geringfügig höher als die Korrelationen der Testhälften und wir können die Nullhypothese nicht zurückweisen, dass beide Tests auf das gleiche Universum möglicher Items zugreifen und Zufallsauswahlen aus einem gemeinsamen Universum möglicher Items sind.

Die niedrigen Korrelationen mit dem Test „Naturwissenschaftliches Denken“ waren zu erwarten, denn dieses Modul ist auf logisches Denken und andere Intelligenzfunktionen ausgerichtet, der HAM-Nat dagegen auf positives Wissen und dessen Anwendung.

Obwohl sich die beiden Testversionen lediglich hinsichtlich der Anzahl der gelösten Items signifikant unterscheiden, deuten die Ergebnisse darauf hin, dass es schwierig ist, parallele Testversionen für naturwissenschaftliches Wissen zu erstellen. Die irrtümliche Annahme einer Äquivalenz (ß-Fehler) in dieser Phase der Testentwicklung wäre nachteiliger als ein Irrtum in die andere Richtung.

Da es trotz vieler Maßnahmen zur Geheimhaltung der Fragen schwer ist zu verhindern, dass Testfragen an die Öffentlichkeit gelangen, müssen für jeden Jahrgang neue Items erzeugt werden. Doch ein gewisser Anteil alter Items mit günstigen Charakteristika sollte wiederverwendet werden, um die Testqualität zu erhöhen und um die Äquivalenz neuer Testversionen mit älteren einzuschätzen. Je mehr der Item-Pool aus vergangenen Tests anwächst, desto größer kann dieser Anteil sein.

Für die Analyse nachfolgender HAM-Nat Versionen sollen Methoden angewendet werden, die eine stichprobenunabhängige Schätzung der Testeigenschaften ermöglichen. Dafür eignen sich Modelle der Item Response Theorie [18]. Sie ermöglichen den Vergleich über verschiedene Testversionen und Studierendenkohorten hinweg. Hierfür benötigen wir einen Pool validierter Items, dessen Aufbau das Ziel unserer Arbeitsgruppe ist.

Danksagung

Wir danken dem Dekan Prof. U. Koch-Gromus und Herrn Dr. B. Andresen für Anregungen und lebhafte Diskussionen und ihre Unterstützung, sowie Herrn D. Münch-Harrach und Herrn C. Kothe für ihre Unterstützung bei der Datenverarbeitung. Diese Studie wird durch den Förderfonds Lehre des Dekanates der Medizinischen Fakultät Hamburg unterstützt.

Interessenkonflikt

Die Autoren erklären, dass sie keine Interessenskonflikte in Zusammenhang mit diesem Artikel haben.

Literatur

1. Bundesministerium für Bildung und Forschung. Hochschulrahmengesetz. BGBl. 2005;I:3835. Zugänglich unter/available from: http://www.bmbf.de/pub/HRG_20050126.pdf
2. Hansestadt Hamburg. Hochschulzulassungsgesetz Hamburg, HmbGVBl. 2004:515-517. Zugänglich unter/available from: http://www.landesrecht.hamburg.de/jportal/portal/page/bshaprod.psml?showdoccase=1&doc.id=jlr-HSchulZulGHArahmen&st=lr
3. Trost G, Flum F, Fay E, Klieme E, Maichle U, Meyer M, Nauels HU. Evaluation des Tests für Medizinische Studiengänge (TMS): Synopse der Ergebnisse. Bonn: ITB; 1998.
4. Trapmann S, Hell B, Weigand S, Schuler H. Die Validität von Schulnoten zur Vorhersage des Studienerfolgs - eine Metaanalyse. Z Padagog Psychol. 2007;21(1):11-27. DOI: 10.1024/1010-0652.21.1.11
5. Ferguson E, James D, Madeley L. Factors associated with success in medical school: systematic review of the literature. BMJ. 2002;324(7343):952-957. DOI: 10.1136/bmj.324.7343.952
6. McManus IC, Smithers E, Partridge P, Keeling A, Fleming PR. A levels and intelligence as predictors of medical careers in UK doctors: 20 year prospective study. BMJ. 2003;327(7407):139-142. DOI: 10.1136/bmj.327.7407.139

7. Wissenschaftsrat. Empfehlungen zur Reform des Hochschulzugangs. Berlin: Wissenschaftsrat; 2004. Zugänglich unter/available from: http://www.wissenschaftsrat.de/download/archiv/5920-04.pdf
8. Trost G. Test für Medizinische Studiengänge (TMS): Studien zur Evaluation, 20. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung; 1996.
9. Koeller O, Baumert J. Das Abitur - immer noch ein gültiger Indikator für die Studierfähigkeit? Politik Zeitgeschichte. 2002;B26. Zugänglich unter/available from: http://www.bpb.de/publikationen/0P7PYG,0,Das_Abitur_immer_noch_eing%FCltiger_Indikator_f%FCr_die_Studierf%E4higkeit.html
10. Janssen PJ. Vlaanderens toelatingsexamen arts-tandarts: resultaten na 9 jaar werking. Ned Tijdschr Geneeskd. 2006;62:1569-81. DOI: 10.2143/TVG.62.22.5002592
11. Smolle J, Neges H, Macher S, Reibnegger G. Aufnahmeverfahren für das Medizinstudium: Erfahrungen der Medizinischen Universität Graz. GMS Z Med Ausbild. 2007;24(3):Doc141. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2007-24/zma000435.shtml
12. Reibnegger, G; Caluba, HC; Ithaler, D; Manhal, S; Neges, HM; Smolle, J. Progress of medical students after open admission or admission based on knowledge tests. Med Educ. 2010; 44(2): 205-214. DOI: 10.1111/j.1365-2923.2009.03576.x
13. Emery JL, Bell JF. The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. Med Educ. 2009;43(6):557-564. DOI: 10.1111/j.1365-2923.2009.03367.x
14. McManus IC, Ferguson E, Wakeford R, Powis D, James D. Predictive validity of the Biomedidcal Admission Test: An evaluation and case study. Med Teach. 2011;33:53-57. DOI: 10.3109/0142159X.2010.525267
15. Hampe W, Klusmann D, Buhk H, Muench-Harrach D, Harendza S. Reduzierbarkeit der Abbrecherquote im Humanmedizinstudium durch das Hamburger Auswahlverfahren für Medizinische Studiengaenge - Naturwissenschaftsteil (HAM-Nat). GMS Z Med Ausbild. 2008;25(2):Doc82. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2008-25/zma000566.shtml.
16. PASW. Predictive Analysis SoftWare. Rel. 18.0.0 ed. Chicago: SPSS Inc.; 2009.
17. Müller KH. Beitrag zum Prüfen der Differenz zwischen 2 Korrelationskoeffizienten. Biometr Z. 1971;13(5):342–361. DOI: 10.1002/bimj.19710130507
18. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, N.J.: L. Erlbaum Associates; 2000.

Korrespondenzadresse:
Prof. Dr. Wolfgang Hampe
Universitätsklinikum Hamburg-Eppendorf, Institut für Biochemie und molekulare Zellbiologie, Martinistraße 52, 20246 Hamburg, Deutschland, Tel.: +49 (0)40/7410-59967, Fax: +49 (0)40/7410-54592
hampe@uke.uni-hamburg.de

Bitte zitieren als
Hissbach J, Klusmann D, Hampe W. Reliabilität des Hamburger Auswahlverfahrens für Medizinische Studiengänge, Naturwissenschaftsteil (HAM-Nat). GMS Z Med Ausbild. 2011;28(3):Doc44.
DOI: 10.3205/zma000756, URN: urn:nbn:de:0183-zma0007562

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2011-28/zma000756.shtml

Eingereicht: *08.10.2010*
Überarbeitet: *29.03.2011*
Angenommen: *01.06.2011*
Veröffentlicht: *08.08.2011*

Reliability of a science admission test (HAM-Nat) at Hamburg medical school

Abstract

Objective: The University Hospital in Hamburg (UKE) started to develop a test of knowledge in natural sciences for admission to medical school in 2005 (Hamburger Auswahlverfahren für Medizinische Studiengänge, Naturwissenschaftsteil, HAM-Nat). This study is a step towards establishing the HAM-Nat. We are investigating

1. parallel forms reliability,
2. the effect of a crash course in chemistry on test results, and
3. correlations of HAM-Nat test results with a test of scientific reasoning (similar to a subtest of the "Test for Medical Studies", TMS).

Methods: 316 first-year students participated in the study in 2007. They completed different versions of the HAM-Nat test which consisted of items that had already been used (HN2006) and new items (HN2007). Four weeks later half of the participants were tested on the HN2007 version of the HAM-Nat again, while the other half completed the test of scientific reasoning. Within this four week interval students were offered a five day chemistry course.
Results: Parallel forms reliability for four different test versions ranged from r_{tt}=.53 to r_{tt}=.67. The retest reliabilities of the HN2007 halves were r_{tt}=.54 and r_{tt} =.61. Correlations of the two HAM-Nat versions with the test of scientific reasoning were r=.34 und r=.21. The crash course in chemistry had no effect on HAM-Nat scores.
Conclusions: The results suggest that further versions of the test of natural sciences will not easily conform to the standards of internal consistency, parallel-forms reliability and retest reliability. Much care has to be taken in order to assemble items which could be used interchangeably for the construction of new test versions. The test of scientific reasoning and the HAM-Nat are tapping different constructs. Participation in a chemistry course did not improve students' achievement, probably because the content of the course was not coordinated with the test and many students lacked of motivation to do well in the second test.

Keywords: Student selection medical school, External validity, Reliability, Admission test

Johanna Hissbach[1]
Dietrich Klusmann[2]
Wolfgang Hampe[1]

1 Universitätsklinikum Hamburg-Eppendorf, Institut für Biochemie und molekulare Zellbiologie, Hamburg, Deutschland

2 Universitätsklinikum Hamburg-Eppendorf, Institut und Poliklinik für Medizinische Psychologie, Hamburg, Deutschland

Introduction

In 2005 Hamburg Medical School started to develop a test of natural sciences (HAM-Nat) as a tool for student admission after a change in federal law allowed German medical schools to select 60% of their student body by admission procedures such as written tests [1], [2].
Until 2008 the Medical Faculty of Hamburg selected candidates solely on the basis of school grade point average (GPA). This is a straightforward approach and GPA is predictive of study success. For the 1986 and 1987 cohorts of medical students Trost et al. [3] reported a correlation of r=.48 between GPA and grades in the written part of the first clinical examination, the correlation for the oral part was r=.34 [3]. In a meta-analysis Trapmann et al [4] report a corrected predictive power of r=.58 for grades in the first section of study. High predictive validity of GPA is also reported in international studies [5] and degree programs other than medicine [4]. In a prospective British study, A-levels showed predictive power for professionalism in the medical field [6].
Nevertheless, GPA as a selective tool is criticized on many accounts [4]:

1. different standards between schools and federal states make GPA scores incomparable;
2. reliability and validity of GPA are insufficient;
3. there are different standards between teachers and classrooms;

4. predictive power for study success later in the curriculum is weak.

Since a large number of candidates apply for medical school, a high level of GPA is necessary for admission. In the years of 2005-2007 applicants for Hamburg Medical School needed GPA scores of at least 1.6 to 1.7 (a low score means high achievement). As GPA is influenced by the type of school, combinations of subjects, and evaluative standards, using GPA as the only admission criterion raises issues of fairness [7]. Additional selection criteria may compensate for some of the shortcomings of GPA.
Some German medical schools use the "Test for Medical Studies" (TMS) to complement GPA in selection. This test was mandatory for all applicants to medical school in the years between 1986 and 1996. It includes questions from the field of natural sciences; however, the targeted construct is not knowledge but the ability to study successfully [8]. Correlations of TMS scores and GPA range from r=.37 to r=.48 and the authors conclude that GPA and TMS measure sufficiently separable facets of academic achievement [3]. Predictive power of the TMS is mainly based on four subtests for abilities required in the medical curriculum (medical and scientific comprehension, quantitative and formal problems, text comprehension, diagrams and tables) [3].
Various countries employ tests of subject specific knowledge relevant for the respective courses [9]. Knowledge tests are used for medical school selection in Belgium [10] and Austria [11]. Reibnegger et al. [12] reported an increase of successful students from 23% to 84% after a demanding admission procedure had been introduced at the university of Graz, Austria (mean percentages of three years before and after the admission procedure had been established). Simultaneously the drop-out rate in the first year of medical school decreased from 10% to 1%. The majority of test items were natural science problems similar to HAM-Nat items.
Since 2003 some British medical schools have introduced the Biomedical Admissions Test (BMAT) for student selection. The subtest "scientific knowledge and application" was a useful predictor for examination marks in the first and second year of study [13]. Predictive power of the second part of the BMAT which entails multiple choice items on problem solving, text comprehension, and the interpretation of tables and figures ("aptitude and skill") is considerably lower [14]. In Germany the HAM-Nat is the only testing program for medical school selection focusing specifically on natural sciences.
With the HAM-Nat test a second selection criterion in addition to GPA will be introduced, a criterion that is uniform for all applicants and might be evaluated consecutively. The HAM-Nat is expected to measure knowledge of natural sciences that is relevant for success in the first two years of the curriculum and thereby help to select applicants with good chance to complete the course successfully. Moreover, an excellent HAM-Nat score may compensate for a low GPA score. The internet page of Hamburg Medical School (http://www.uke.uni-hamburg.de/studienbewerber) not only offers information about the curriculum but additionally exhibits HAM-Nat test items for a self-test of knowledge in natural sciences. Potential applicants may examine their motivation to study medicine and assess their chances to succeed. Hamburg Medical School deliberately aims at pre-selection by self-evaluation. Preparation for the HAM-Nat is tantamount to a preparation for the first two years of study because the HAM-Nat examines basic knowledge required for the science classes during the first two years of study.
A preliminary version of the HAM-Nat was presented 2006 to a sample of high school students. From this pilot study a first 2006 version was derived and tested with the 2006 cohort of already admitted students [15]. Subsequently, new items were generated for a 2007 version of the test. The existence of two test versions raises the question of test equivalence.
This study attempts to answer this question and will additionally examine retest reliability. We also analyze the effect of a five-day crash course (training in basic chemistry) on HAM-Nat scores and the relation of the HAM-Nat to a test of "scientific reasoning". This test resembles the TMS subtest "medical and scientific comprehension".

Methods

Hampe et al. [15] describe the test development of the 2006 HAM-Nat test form (HN2006). After exclusion of 8 items with low item-total correlations, the HN2006 consists of 52 multiple choice items from mathematics, chemistry, physics, and biology. To create a parallel test form for 2007 (HN2007), high school teachers and university lecturers from clinical and basic science departments generated 60 new items similar in content and structure to the HN2006 items.
Example for a HAM-Nat-question:

Oxidation of an aldehyde yields...
A) an ester.
B) a ketone.
C) a carboxylic acid.
D) an alcohol.
E) an alkene.

Each item presents one correct answer and four distractors, testees have 1.5 minutes to answer each question. The topics covered in the test and some sample items from HN2006 and HN2007 are published on the internet page of Hamburg medical school for self-testing (http://www.uke.uni-hamburg.de/studienbewerber).

Test of "scientific reasoning"

The "scientific reasoning" test is similar to the TMS subtest "medical and scientific comprehension" with regard to form and content. Both tests comprise 24 multiple choice items and both were developed by ITB-Consulting. Each question starts with the description of a scientific problem. Subsequently, the testee has to decide which

statement out of five options following the text is true. The duration of the test is limited to 55 minutes. Prior scientific knowledge is not needed to answer these questions since the test is designed to measure intellectual abilities relevant to the medical curriculum: comprehension of complex problems and deductive reasoning. Rights to use the test were purchased from ITB-Consulting.

Chemistry course

The chemistry department of Hamburg University regularly offers this five-day course to first year medical students before the beginning of the term. The intention of this optional course is to level previous knowledge of students. Several parallel courses of 30-40 students are run by tutors. The tutors present topics from senior years of secondary school, e.g. the concept of matter, chemical reaction, and the structure of organic compounds, and afterwards students work on problems. The course's contents are similar to HAM-Nat topics. However, tutors did not teach to the test as they were not familiar with the HAM-Nat.

Study design

First testing: Parallel forms

The HN2006 was divided into two halves A and B of 26 items each. As the test had been accessible on the internet, participants might have known these items. The HN2007 comprises two halves C and D of 30 new items. Each participant worked on two halves from each test version (AC, AD, BC, or BD), namely 26 old items from HN2006 and 30 new items from HN2007 (see Figure 1). Before the second testing, each student had the opportunity to attend the five day training course. Participants stated how many days of the course they attended.

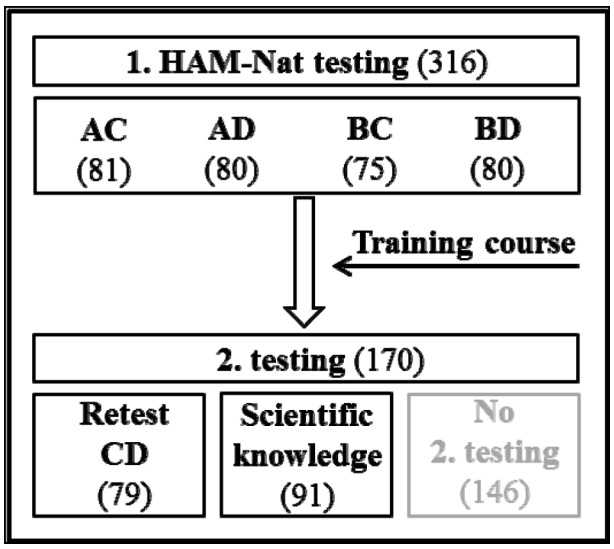

Figure 1: Study design and number of participants

Second testing: Retest and test of "scientific reasoning"

Participants were randomly assigned to two groups. Four weeks after the first testing 96 participants took the complete HN2007, meaning test halves C and D. Therefore, they had already answered one half of the items while the other half was unknown. The following week, the second group of participants (N=91) took the subtest "scientific reasoning". The test was conducted by our study group and invigilated by members of faculty. The "scientific reasoning" test was specifically conducted for this study, irrespective of the official, nationwide testing of the TMS.

Participants

All students of the 2007 cohort were offered study participation in the first semester orientation week. Study participation was voluntary, and all students gave written informed consent. 316 students (77% of the cohort) agreed to participate (see Figure 1). One third of the sample was male, two thirds female which corresponds to the distribution in the total cohort. The mean secondary school GPA was 1.8. The second test was conducted in a compulsory class during the first term. As opposed to the participation in the orientation week, many attendees were not willing to retake the test. Results of the first and second testing could be matched for 170 students (54% of the original sample). No significant differences regarding GPA and gender distribution were found between the group of test repeaters and those who denied the retest. 91 participants worked on the "scientific reasoning" test and 79 wrote the HN2007 again. The effect of the chemistry course can be evaluated with data from 52 students, who took the HAM-Nat twice and stated the number of days that they attended the course. 15 students stated they had attended 3 or fewer days, while 37 attended more than 3 days.

Statistical analysis

Parallel forms reliability requires true values and error variances to be equal. Equal means and distributions of data, as well as a high correlation between test versions, indicate high parallel forms reliability. Retest reliability assumes that between two assessments the participants' true scores are constant as is measurement error. It reflects the degree in which repeated measurement with a certain measure on the same population reveals according results.

Pearson correlations are calculated to quantify parallel forms reliability and the correspondence of HAM-Nat and "scientific reasoning". Retest reliability of HN2007 was assessed by means of Spearman's rank correlation coefficient.

Cronbach's α is the expected value of a correlation of two randomly selected item sets (of k items) from the universe

of all possible items for the measured construct. If HN2006 and HN2007 are parallel forms, the correlations of test halves must be as high as their internal consistencies. If the correlations are different in size, the item set is either drawn from different item universes or items are not randomly selected.

A general linear model (GLM) of HN2006 and HN2007 total scores was employed, with "test version" as a repeated measurement factor (within subjects factor) and "group" (AC, AD, BC, BD) as a between subjects factor. A significant repeated measurement factor means that test versions differ in difficulty, while significant interaction effects with group point to differences in test halves within one test version.

To estimate the effect of the chemistry course on chemistry test item performance, the variable "attendance at the course" was dichotomized (0-3 vs. >3 days) and included in a new model as a between subjects factor as well as the within subject factors "items separated by subject" (chemistry vs. other questions) and "time" (first testing vs. retest). PASW 18 for Windows [16] was used for these analyses.

Results

Internal consistency and parallel forms reliability

Inter item correlations for all scales ranged from r=-.22 to r=.53 (mean: r=.06), internal consistencies from α=.56 to α=.69 (see Table 1), and parallel forms correlations from r=.53 to r=.67 (see Table 2).

Table 1: Statistics of test scales and test versions

Test version	HN2006		HN2007	
	A	B	C	D
Valid N	162	155	157	160
Number of items	26	26	30	30
Cronbach's alpha (KR-20)	.687	.630	.628	.557
Mean item difficulty	.440	.451	.405	.351
Mean item variance	.206	.204	.204	.199
Mean inter-item-correlation	.076	.060	.052	.037
Mean item-total-correlation	.229	.195	.181	.147
Scale mean	11.45	11.73	12.14	10.53
Scale variance	15.72	13.42	15.58	12.94
Correlation with GPA	-.278	-.126	-.221	-.335

Table 2: Correlation between different test halves

		HN 2007	
		Test half C	Test half D
HN 2006	Test half A	.674	.660
	Test half B	.531	.532

Retest reliability

Retest reliability was calculated for HN2007. Pearson's rank correlation for test half C was r_{tt}=.52 (n=46), for test half D r_{tt}=.61 (n=34) (see Figure 2). The corresponding Pearson correlations were r_{tt}=.54 and r_{tt}=.56. Some participants scored considerably worse in the retest as compared to the first testing. Exclusion of 9 participants with retest scores below 6 did not raise the correlation coefficient (test half C r_{tt}=.45, n=39; test half D r_{tt}=.61, n=32), even though Figure 2 might suggest such an effect.

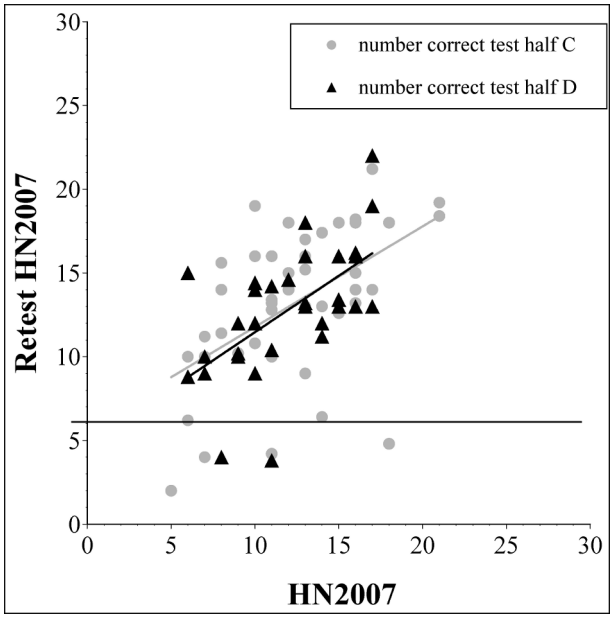

Figure 2: Test/Retest-Correlations for test halves C and D (HN2007) and their respective regression lines. Below the horizontal line are those participants scoring lower than chance level in the second testing.

Differences between test versions HN2006 and HN2007

The general linear model (GLM) gives a more detailed look at differences between test versions. A GLM with the factors "test version" (HN2006 vs. HN2007) as a repeated measurement factor and test half (A or B vs. C or D) as a between subjects factor showed that significantly

fewer HN2007 items were solved correctly as compared to the HN2006 version (38.5% vs. 45.2%, $F_{1,312}$=101.5; p<.001). While all participants scored equally high in both test halves of version HN2006 ($F_{1,312}$=2.3; p=.128), test half D was more difficult than test half C in the HN2007 version (35.1% vs. 40.6% correct answers, $F_{1,312}$=11.4; p=.001). Gender, included as a between subjects factor, had no significant effect on performance in the different test versions (F=.468, p=.495), even though males showed higher total scores as compared to females (44% vs. 40% correct answers, T=-2.64; p=.009).

Effect of the chemistry course

Scores of the first and second testing were analyzed separately for chemistry and other items (biology, mathematics, and physics) to check the effect of the chemistry course on performance. Fewer chemistry items were answered correctly as compared to items from the other subjects (35.8% vs. 43.4% correct answers, $F_{78,1}$=25.6, p<.001). There was neither an improvement nor a decline of HN2007 test results after the course, not even chemistry items were answered correctly more often (interaction effect: $F_{1,78}$=0.26; p=.610). The dichotomized variable "days of participation in the course" (0-3 vs. 4-5 days of attendance) showed no significant effect on HN2007 total scores ($F_{1,50}$=2.4; p=.124) or chemistry scores ($F_{1,50}$=0.1; p=.759). The sample for this analysis is reduced to n=52. Including gender in the model yielded no significant interaction effects (all p>.289).

Publicity of test items

In the retest condition, participants had already seen half of the HN2007 items, while the other half was new. Known items were not significantly more often answered correctly as compared to the first test (41.5 % 40.1 %; $F_{1,50}$=0.4; p=.543). Exclusion of participants with very low retest scores did not alter results.

Correlation of HAM-Nat and GPA

Correlation coefficients for HAM-Nat and GPA scores ranged between r=-.34 and r=-.13 (see Table 1) for the different versions (mean correlation r=-.24). GPA and the test "scientific reasoning" showed a correlation of r=-.11 (n=90).

Correlation of the HAM-Nat and "scientific reasoning"

The correlation of the subtest "scientific reasoning" and HN2006 test halves A and B were r=.34. Correlation coefficients for HN2007 test halves C and D were r=.19 and r=.23, respectively (see Figure 3). For the combined test halves HN2006 (A+B) and HN2007 (C+D) the correlation coefficients were r=.34 and r=.21. The two coefficients did not differ significantly (p=.350; test with Fisher's z [17]).

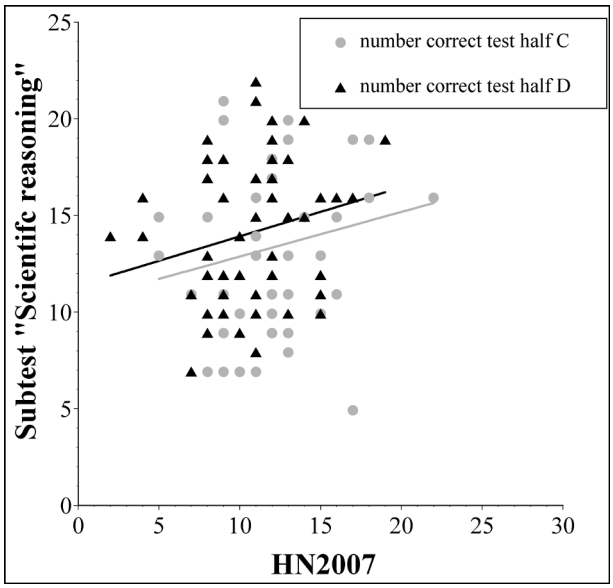

Figure 3: Correlations of the subtest „scientific reasoning" with the HN2007 test Halves C and D and their respective regression lines.

Discussion

Results can be summarized as follows:

1. Significantly more old HN2006 items were solved correctly as compared to new HN2007 items, and taking HN2007 twice did not improve test performance.
2. HN2006 and HN2007 did neither differ with regard to their internal consistencies nor with regard to their correlations with a third test "scientific reasoning".
3. Internal consistencies of the different test versions were not significantly different from the correlation between test halves (parallel forms reliability).

Why is the HN2006 easier than the HN2007? Maybe some participants were familiar with the old test items due to their publication on the internet. However, we assume that participants did not prepare for the test because they had already been admitted to medical school and nothing was at stake. We do not know how many students took the internet self-test. However, taking the HN2007 twice within a four week period did not lead to better results. Why should the supposedly infrequent visit of this internet page have an effect? It is more likely that test developers produced more difficult items.

On the one hand, varying difficulties of HAM-Nat test forms are not problematic since the purpose of the test is to rank applicants in a combined score of HAM-Nat and further admission criteria (GPA, further tests). As long as tests produce the same rank ordering, they are exchangeable. However, a test used for student selection should exhibit a profile which is constant over different cohorts. Rank correlation coefficients are a measure of reproducibility. For test halves C and D of the HN2007 they were r=.52 and r=.61. These are not very high values given that participants had seen the same items four weeks

prior to the second testing. This low level of reproducibility might be due to an important source of error which applies to the whole study design: since stakes were low, test score variation is not only due to differences in knowledge but also to differences in test motivation. This is especially true for the retest condition with just above half of the sample taking part. At this time point, participants were busy with the first weeks of the term. Therefore, the low retest correlation is probably an underestimation.

The especially low performance in chemistry items could be explained by the fact that most German schools introduce chemistry classes later into the curriculum than other natural sciences. Moreover, more students drop this subject in sixth form as compared to other science classes. If, for example, biology is dropped in sixth form, pupils still have had more years studying biology as compared to the scenario where chemistry classes are dropped. Offering a training course in chemistry seems worthwhile. But why did we not see better results in the chemistry items of the HAM-Nat retest? For this part motivation to do well is very important and probably participants were not motivated enough. Another explanation could be that HAM-Nat items covered knowledge which was not taught in the course. This finding draws attention to the process of writing items. New items should correspond to the typical teaching material that applicants use for test preparation. Only if this is the case, test preparation can improve chances to be admitted – one of the intended effects of the HAM-Nat. To improve further versions of the HAM-Nat test, a list of topics was published in 2008 to help applicants with their preparation. All subsequent HAM-Nat items can be reliably assigned to one or more topics of the list of subjects.

HN2006 items had been preselected by item total correlation in a first test run which was not the case for HN2007. This might explain the slightly smaller – yet insignificant – internal consistency of HN2007. To check this, we excluded items with corrected item total correlations <.10 from both scales and recalculated internal consistencies. HN2006 contained merely 5 items below .10 while HN2007 contained 15 items that had to be excluded. After exclusion of these items, internal consistencies for all test halves amounted up to values between .60 and .70. Therefore, internal consistencies are only slightly higher than correlations of test halves, and we cannot reject the null hypothesis that both tests are drawn from the same item universe and that they are randomly selected from this universe.

We expected correlations of the HAM-Nat and the external criterion “scientific reasoning” to be low as the “scientific reasoning” is targeted on ability to reason and intelligence while the HAM-Nat test is targeted on knowledge and application of knowledge.

Even though the test versions HN2006 and HN2007 only differ with regard to the number of correctly solved items, results indicate that it is difficult to develop parallel test forms for knowledge of sciences. Erroneously assuming that test forms are parallel (beta error) at this stage of test development is more harmful than the contrary error. Despite many actions to prevent that items are made public, new items have to be written every year. However, a certain proportion of old items with good psychometric properties should be reused to raise test quality and to estimate equivalence of new test versions. The larger the item pool, the more items can be reused. Methods that are able to estimate sample independent test characteristics should be used for subsequent HAM-Nat test versions. Models within the item response theory (IRT) framework [18] allow comparisons across different test versions and cohorts of students. Therefore, the aim of our project is to assemble a pool of validated items.

Acknowledgements

We are grateful to Prof. U. Koch-Gromus and Dr. B Andresen for open discussions and their collaboration, and we would like to thank D. Münch-Harrach and C. Kothe for their support in data management. This research was funded by the “Foerderfonds Lehre”, a grant of the Universitaetsklinikum Hamburg Eppendorf.

Competing interests

The authors declare that they have no competing interests.

References

1. Bundesministerium für Bildung und Forschung. Hochschulrahmengesetz. BGBl. 2005;I:3835. Zugänglich unter/available from: http://www.bmbf.de/pub/HRG_20050126.pdf
2. Hansestadt Hamburg. Hochschulzulassungsgesetz Hamburg, HmbGVBl. 2004:515-517. Zugänglich unter/available from: http://www.landesrecht.hamburg.de/jportal/portal/page/bshaprod.psml?showdoccase=1&doc.id=jlr-HSchulZulGHArahmen&st=lr
3. Trost G, Flum F, Fay E, Klieme E, Maichle U, Meyer M, Nauels HU. Evaluation des Tests für Medizinische Studiengänge (TMS): Synopse der Ergebnisse. Bonn: ITB; 1998.
4. Trapmann S, Hell B, Weigand S, Schuler H. Die Validität von Schulnoten zur Vorhersage des Studienerfolgs - eine Metaanalyse. Z Padagog Psychol. 2007;21(1):11-27. DOI: 10.1024/1010-0652.21.1.11
5. Ferguson E, James D, Madeley L. Factors associated with success in medical school: systematic review of the literature. BMJ. 2002;324(7343):952-957. DOI: 10.1136/bmj.324.7343.952
6. McManus IC, Smithers E, Partridge P, Keeling A, Fleming PR. A levels and intelligence as predictors of medical careers in UK doctors: 20 year prospective study. BMJ. 2003;327(7407):139-142. DOI: 10.1136/bmj.327.7407.139
7. Wissenschaftsrat. Empfehlungen zur Reform des Hochschulzugangs. Berlin: Wissenschaftsrat; 2004. Zugänglich unter/available from: http://www.wissenschaftsrat.de/download/archiv/5920-04.pdf

8. Trost G. Test für Medizinische Studiengänge (TMS): Studien zur Evaluation, 20. Arbeitsbericht. Bonn: Institut für Test- und Begabungsforschung; 1996.

9. Koeller O, Baumert J. Das Abitur - immer noch ein gültiger Indikator für die Studierfähigkeit? Politik Zeitgeschichte. 2002;B26. Zugänglich unter/available from: http://www.bpb.de/publikationen/0P7PYG,0,Das_Abitur_immer_noch_eing%FCltiger_Indikator_f%FCr_die_Studierf%E4higkeit.html

10. Janssen PJ. Vlaanderens toelatingsexamen arts-tandarts: resultaten na 9 jaar werking. Ned Tijdschr Geneeskd. 2006;62:1569-81. DOI: 10.2143/TVG.62.22.5002592

11. Smolle J, Neges H, Macher S, Reibnegger G. Aufnahmeverfahren für das Medizinstudium: Erfahrungen der Medizinischen Universität Graz. GMS Z Med Ausbild. 2007;24(3):Doc141. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2007-24/zma000435.shtml

12. Reibnegger, G; Caluba, HC; Ithaler, D; Manhal, S; Neges, HM; Smolle, J. Progress of medical students after open admission or admission based on knowledge tests. Med Educ. 2010; 44(2): 205-214. DOI: 10.1111/j.1365-2923.2009.03576.x

13. Emery JL, Bell JF. The predictive validity of the BioMedical Admissions Test for pre-clinical examination performance. Med Educ. 2009;43(6):557-564. DOI: 10.1111/j.1365-2923.2009.03367.x

14. McManus IC, Ferguson E, Wakeford R, Powis D, James D. Predictive validity of the Biomedidcal Admission Test: An evaluation and case study. Med Teach. 2011;33:53-57. DOI: 10.3109/0142159X.2010.525267

15. Hampe W, Klusmann D, Buhk H, Muench-Harrach D, Harendza S. Reduzierbarkeit der Abbrecherquote im Humanmedizinstudium durch das Hamburger Auswahlverfahren für Medizinische Studiengaenge - Naturwissenschaftsteil (HAM-Nat). GMS Z Med Ausbild. 2008;25(2):Doc82. Zugänglich unter/available from: http://www.egms.de/static/de/journals/zma/2008-25/zma000566.shtml.

16. PASW. Predictive Analysis SoftWare. Rel. 18.0.0 ed. Chicago: SPSS Inc.; 2009.

17. Müller KH. Beitrag zum Prüfen der Differenz zwischen 2 Korrelationskoeffizienten. Biometr Z. 1971;13(5):342–361. DOI: 10.1002/bimj.19710130507

18. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, N.J.: L. Erlbaum Associates; 2000.

Corresponding author:
Prof. Dr. Wolfgang Hampe
Universitätsklinikum Hamburg-Eppendorf, Institut für Biochemie und molekulare Zellbiologie, Martinistraße 52, 20246 Hamburg, Deutschland, Tel.: +49 (0)40/7410-59967, Fax: +49 (0)40/7410-54592
hampe@uke.uni-hamburg.de

Please cite as
Hissbach J, Klusmann D, Hampe W. Reliability of a science admission test (HAM-Nat) at Hamburg medical school. GMS Z Med Ausbild. 2011;28(3):Doc44.
DOI: 10.3205/zma000756, URN: urn:nbn:de:0183-zma0007562

This article is freely available from
http://www.egms.de/en/journals/zma/2011-28/zma000756.shtml

Received: *2010-10-08*
Revised: *2011-03-29*
Accepted: *2011-06-01*
Published: *2011-08-08*