Zentrum für Testentwicklung und Diagnostik
am Psychologischen Institut der Universität Fribourg

Eignungsdiagnostik und Medizinstudium

Symposiumsbericht herausgegeben von
K.-D. Hänsgen, R. Hofer und D. Ruefli

Bericht 2 (1996)

Inhaltsverzeichnis

©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

Eignungstests und Medizinstudium

Am 27. und 28 Oktober 1995 trafen sich in Fribourg Psychologen, Pädagogen, Mediziner und Vertreter anderer Disziplinen, um ausführlich alle wissenschaftlichen Fragen zu diskutieren, die mit einem möglichen **Eignungstest für das Medizinstudium in der Schweiz** verknüpft sind.

Die Veranstaltung wurde gemeinsam mit dem Psychologischen Institut der Universität Fribourg unter der Schirmherrschaft der Schweizerischen Hochschulkonferenz durchgeführt. Wir danken Herrn Prof. Perrez, dem Direktor des Psychologischen Instituts, für seine Unterstützung, das Anliegen der Veranstaltung umzusetzen, welches darin bestand, eine Bilanz hinsichtlich der **wissenschaftlichen Absicherung** von Studieneignungstests zu ziehen und **Anforderungen** für eine mögliche Anwendung zu definieren. Dem Plenum am Freitag mit sehr angeregten Diskussionen folgte ein Workshop mit den internationalen Gästen am Samstag.

Im Mittelpunkt stand die Diskussion **internationaler Erfahrungen** mit Studienzulassungstests, welche in der Mehrzahl der Industrieländer eingesetzt werden und dort - unter den Bedingungen der Notwendigkeit eines Numerus Clausus - als wissenschaftlich fundierte, faire und praktikable Methode der Regulierung des Studienzuganges anerkannt sind.

R. Hambleton (University of Massachusetts, USA) stellte den Medical College Admission Test (MCAT) vor, der von der Association of American Medical Colleges als Eingangstest verwendet wird. Es handelt sich um einen Studierfähigkeitstest, der vor allem die Problemlösefähigkeit prüft. Er wird in regelmässigen Abständen an die Studienanforderungen angepasst und revidiert. Ein zweiter Schwerpunkt war die Darstellung von Kriterien für Übersetzungen von Test- und Prüfungsfragen in andere Sprachen. Hierzu liegen international anerkannte Regeln der International Test Commission vor, die auch für die Übersetzungen in der Schweiz wertvolle Anregungen geben. ***M. Beller*** (University Tel Aviv, Israel) berichtete über den Eignungstest, der in Israel für die Zulassung zum Medizinstudium verwendet wird. Bemerkenswert ist, dass dieser Test in sechs Sprachen absolviert werden kann. Die Äquivalenz der jeweiligen Übersetzungen wird laufend überprüft. Es wurden praktische Kontroll- und Korrekturmöglichkeiten vorgestellt, die eine Chancengleichheit gewährleisten. ***Ch. Stage*** (Universität Umeå, Schweden) demonstrierte die Prinzipien des Swedish Scholastic Aptitude Test (SweSAT), der seit 1977 eingesetzt wird und ebenfalls 1991 modifiziert worden ist. In sechs Untertests werden für das Studium wichtige Leistungen geprüft. Dieser Test wird zweimal im Jahr durchgeführt und prognostiziert ebenfalls die Studieneignung.

Besonderer Wert wird auf einen laufenden Abgleich mit den Studienanforderungen bei der Entwicklung gelegt. ***G. Trost und E. Klieme*** (Institut für Test- und Begabungsforschung Bonn, Deutschland) konnten neueste Ergebnisse zum in Deutschland verwendeten Test für Medizinische Studiengänge (TMS) referieren, welche seine sehr hohe prognostische Validität für den Studienerfolg unter Beweis stellen, die einer Verwendung der einfachen Abiturnote als Kriterium überlegen ist. Zum Problem der Testfairness konnte anhand empirischer Daten gezeigt werden, dass mit einer Verwendung des Tests die Bedingungen für deren Einhaltung gegeben sind.

Bei allen vorgestellten Testentwicklungen wird grosser Wert darauf gelegt, dass kein medizinspezifisches Wissen für die Lösung der Aufgaben notwendig ist, um die Trainierbarkeit so gering wie möglich zu halten. Dies ist ein wichtiges Merkmal für die soziale Verträglichkeit von Zulassungstests, damit teure Vorbereitungskurse wirkungslos und damit unnötig bleiben.

R. Bloch (Universität Bern) berichtete über Erfahrungen mit einem komplexen Auswahlverfahren an der McMaster-Universität in Kanada, wo vor allem Eignungsgespräche eingesetzt werden. ***F. Baumann*** (Universität Genf) diskutierte den Stellenwert von Wissen und Fähigkeiten, die ein kompetenter Arzt heute benötigt. ***U. Schallberger*** (Universität Zürich) konnte zeigen, dass Nutzen und Fairness die beiden wichtigsten Bewertungskonzepte für Selektionsverfahren sind. Anhand theoretischer Überlegungen demonstrierte er Probleme und mögliche Lösungen für die Gewährleistung von Chancengleichheit für verschiedene Gruppen.

In einem Beitrag des Zentrums für Testentwicklung ***(R. Hofer, D. Ruefli und K.-D. Hänsgen)*** wurden erste Ergebnisse eines **Probelaufs des Eignungstests** an einer deutsch- und französischsprachigen Stichprobe mit den jeweiligen Sprachformen (durchgeführt am Collège Sainte-Croix Fribourg) vorgestellt. Trotz fehlender Bewerbungsmotivation erreichte der Test annähernd gleiche Gütekriterien wie in Deutschland. Der Test differenziert bezüglich der Leistung und damit der Studieneignung in ausreichendem Masse, so dass eine Studienzulassung mit diesem Verfahren gerechtfertigt wäre. Wie in Deutschland werden im Mittel etwa 50% aller Aufgaben richtig gelöst.

Schussfolgerungen des Symposiums für eine mögliche Anwendung eines Eignungstests in der Schweiz fasste ***N. Ischi*** (Generalsekretär der Schweizerischen Hochschulkonferenz) zusammen.

In diesem Band werden ausgewählte Beiträge für die weitere Diskussion über den Nutzen und mögliche Probleme der Anwendung eines Eignungstests in der Schweiz veröffentlicht. Dies geschieht auch in der Hoffnung, die Diskussion weiter zu versachlichen und eine realistische Einschätzung des Tests zu ermöglichen. Wir glauben, dass letztlich zur Anwendung eines Tests keine gleichwertige Alternative besteht, wenn ein Numerus Clausus notwendig wird.

Eignungsgespräche sind nicht bezahlbar, wenn sie mit allen Bewerberinnen und Bewerbern durchgeführt werden sollen. Sie wären allenfalls als Korrektiv für den Personenkreis im „Grenzbereich“ der Zulassung anstatt einer Warteliste sinnvoll. Ein Praktikum bietet, wenn Plätze überhaupt in der geforderten Menge zur Verfügung gestellt werden könnten, keine Möglichkeit zur Kapazitätsregelung, da erfahrungsgemäss allenfalls eine sehr kleine Zahl von Studierwilligen aufgrund der Eindrücke im Praktikum von der Bewerbung abgehalten wird. Maturitätsnoten sind in der Schweiz nicht vergleichbar. Der Test ist bezüglich seiner Prognosekraft für Studienerfolg wissenschaftlich überprüft. Es liegen positive Erfahrungen mit einem solchen Instrument aus einer beträchtlichen Zahl von Ländern vor. Er bleibt bezahlbar und ist fair gegenüber den Sprachgruppen und den Geschlechtern.

Klaus-D. Hänsgen

Direktor des ZTD

Weitere Literatur:

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

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

Hofer, R., Ruefli, D., Hänsgen, K.-D.(1995). Der Eignungstest für das Medizinstudium in der Schweiz. Ein Probelauf. Berichte des ZTD Nr.1

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

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

Experiences with the Swedish Scholastic Aptitude Test

Christina Stage

Umeå University, Department of Educational Measurement

The Swedish School System

Sweden has a long tradition of compulsory, comprehensive education. As early as in 1842 the first law was passed and signed by the king that there should be at least one proper school with a trained teacher in each municipality in the country. Since then there have been several school reforms which have tried to create a school system combining quality with equality.

Primary and secondary education are common for all children in Sweden. In Autumn 1995, 930 000 pupils attended primary and secondary school.

After nine years of compulsory, integrated education the students can choose between different study lines (now changing to programmes) in upper secondary school. Admittance to, or rather placement in, upper secondary school is based on average marks from the ninth year in lower secondary school. The choices offered are between five different theoretical study lines/programmes, all preparing for higher education, and about 35 different, vocationally oriented study lines/programmes. About half the students choose one of the theoretical study lines.

The Marking System in Sweden

In primary school there is no marking. In lower secondary school marks are given only in grades eight and nine.

Up to now the marking system in Swedish secondary schools has been norm- or group-referenced. The reference has been made to all pupils in the country of the same grade each year. The scale of marks has ranged from one to five, where one has been the lowest and five the highest and three should be the average mark in each subject. The marks have been comparable all over the country and the comparability has been ensured by centrally constructed and administered standardised tests in the core subjects. The results from these tests were used to decide the average of the class, i.e. the individual teacher was told how his/her class compared to other classes in the country. If the class average was above or below the average of all classes in the country the teacher was supposed to adjust the class average accordingly. The test results were not decisive for the marks of individual pupils.

In upper secondary school marks are given at the end of each term. The marks have, up to now, been norm or group-referenced but the norm-groups have been all other students studying the same subject. That means that as different subjects are studied at different study lines the norm-groups have been different for the different study lines.

The marking system in Sweden is now being changed from a norm-referenced to a criterion- or goal-referenced system. The situation is a bit confused at the moment as the goals or criteria for different marks have not yet been finally decided.

At the university level the marking is criterion-referenced with only three levels: failed, passed and passed with distinction.

The Swedish Scholastic Aptitude Test

The SweSAT was introduced in 1977 in connection with a reform of the universities and colleges. It was felt that an admission test would provide a possible solution to two basic problems (1) how to find a method of selection which could be used for applicants without formal qualifications; and (2) how to reduce the decisive role played by marks in the selection process. When the test was first introduced it was, however, only made available for a relatively small group of applicants (those who were at least 25 years old and had at least four years of work experience). Only since 1991 has the test been used for all applicants.

From 1977 to 1989, as long as the use of the SweSAT was restricted to the above-mentioned group the number of persons taking the test was approximately 10 000 each year; 6 000 in the spring and 4 000 in the autumn. Since 1990 the number of testtakers has increased dramatically to around 140 000 persons each year; 75 - 80 000 in the spring and 55 - 60 000 in the autumn.

At present the test consists of 148 multiple choice questions distributed on six subtests. The results are transformed to a standard scale from 0.0 to 2.0 where 2.0 is the highest result. The test is administered twice a year, in spring and autumn. Students are allowed to take the test as many times as they wish and for those who have several results the best one is used for application. In principle it is optional to take the test; in reality, however, test results can be seen as necessary, since only applicants with top marks dare refrain from taking the test. The content of the test is shown in table 1.

Table 1: The Swedish Scholastic Aptitude Test

Subtest	Abbreviated	Items	Time (min)
Vocabulary	WORD	30	15
Data Sufficiency (Quantitative reasoning)	DS	20	45
Reading comprehension	READ	24	60
Interpretation of diagrams, tables and maps	DTM	20	55
General information	GI	30	25
English reading comprehension	ERC	24	50
Total		148	4 hrs 10 min

Vocabulary (WORD) measures understanding of words and concepts, and consists of items where the task is to identify which of five presented words has the same meaning as a given word. Both Swedish and foreign words are included in the subtest.

Data Sufficiency (DS) aims at measuring numerical reasoning ability. In each item a problem is presented, and the task is to decide whether the information presented is sufficient to allow solution of the problem. The response format is fixed, so each item presents the same five alternatives. The subtest is designed to put as little premium as possible on mathematical knowledge and skills in favour of problem-solving and reasoning.

Reading Comprehension (READ) measures Swedish reading comprehension in a wide sense. The examinees are presented with six texts and four multiple choice questions in relation to each text. Each text comprises about one printed page. Some items ask about particular pieces of information but most items are designed to require understanding of larger parts of the text or the text in its entirety.

Interpretation of Diagrams, Tables and Maps (DTM) consists of 10 collections of tables, diagrams and/or maps which present information about a topic, with two multiple choice questions in relation to each collection. The degree of complexity of the items varies from simply reading off a presented graph, to some where information from different sources must be combined.

General Information (GI) measures knowledge and information from many different areas. The test is broader than traditional school achievement tests and asks about information that a person may acquire over an extended period of time in different contexts such as work and education, or social, cultural and political activities.

English Reading Comprehension (ERC) is of the same general type as the subtest READ. However, in this subtest there is more variability as to both the texts and item formats used. The test consists of 8 to 10 texts of different lengths. Most texts are followed by one or more multiple choice questions with four alternatives. In one of the texts, some words are omitted, and the examinee is supposed to select the omitted word from four alternatives presented alongside the text.

The SweSAT is supposed to measure acquired (developed) abilities and it makes use of the kind of verbal and mathematical skills that develop over the years, both in and out of school. The content of the test does not reflect any specific curriculum although it is designed to be consistent with school based learning.

The test is designed for selection to all different types of university programmes and therefore it is intended to measure the students' general aptitude for studies. Since the test is a selection test it is supposed to rank the applicants as fairly as possible according to their expected academic success. Other requirements on the test are:

- The test should be in line with the aims and content of higher education.
- The test must not have negative effects on the education in upper secondary school.
- It should be possible to score the test fast, cheaply and objectively.
- It should not be possible for an individual to improve his/her test result by means of mechanical exercises or by learning special principles for problem solving.
- The examinees should experience the test as meaningful and suitable.
- The demand for unbiased recruitment should be observed. No group should be discriminated against because of gender or social class.
- The test should also be varied and cover many different content areas. It is possible to find the answers to roughly half of the questions in the material provided. In order to answer the remaining questions some background knowledge is necessary.

On the whole the test has been surprisingly well received by testtakers as well as educational institutions. It is now accepted as a major alternative to school marks as selection instrument and it has even been suggested as a substitute now that the marking system is being changed.

One reason for this acceptance of the SweSAT might be that the test was introduced "as a second chance" and has been regarded as such. Another reason might be that the test along with the scoring key has always been made public as soon as the test has been administered, which means that the test-

takers have the opportunity to control (and discuss) their results on every single item. A final reason might be that the test is a good one or at least that the testtakers really experience it as a meaningful and suitable selection instrument for higher education.

Selection to Higher Education in Sweden

In Sweden there are six universities, 16 university colleges and six specialized institutions of higher education. The difference between the universities and the other institutions of higher education is that graduate programmes are only offered by the universities.

Approximately 50 000 students are admitted to higher education every year and quite a few of the study programmes offered have many more applicants than available study places. As a result of the high unemployment rate, the competition for study places has been growing. Even though the government has increased the number of study places the number of applicants has increased still more. Therefore selection for the study places must take place and for many of the study programmes the competition is very keen.

The selection to higher education has changed substantially during the last three decades. Previously the only selection instrument was marks from upper secondary school. In 1977 the SweSAT was introduced as a selection instrument, but only for a small group of applicants. In 1991 the selection rules were changed again and since then all applicants can use test results as an alternative to marks.

A noteworthy feature of the Swedish selection system is that the applicants may use either marks or test results, whichever is most favourable. This means that, even though it is optional to take the test, so far, most students are taking it. One of the main reasons for making SweSAT scores available for all applicants was to make the average marks from upper secondary less crucial than they had been before and to make it easier for students to be admitted to higher education immediately after leaving upper secondary school. The SweSAT was to give students who had not managed to get top marks, a second chance of admittance.

Originally selection to approximately 60 per cent of the study places was made on the basis of the applicants' marks and selection to the remaining 40 per cent was based on the results on the SweSAT. Since 1993 the universities and colleges are autonomous in deciding their admission procedures and selection devices. No major changes have taken place yet, however, and still usually 60 per cent of the study places are allocated on basis of average marks from upper secondary school and 40 per cent on basis of test scores.

Selection to Medical Education in Sweden

Medical education is provided at the six universities in Sweden and it is one of the study programmes for which the competition is the very hardest. These study programmes have also been those most eager to make use of the right to decide for themselves how to select students, and therefore the systems vary quite a bit.

At Umeå University 61 students are admitted each term and starting this autumn the selection is made in two stages. In stage one, the 122 applicants with the highest scores on the SweSAT (top marks in the core subjects, i.e. the subjects necessary to qualify for the programme, are given some extra credit) are chosen and invited for an interview. The interviews are made by teachers/doctors at the university and the aim is to sort out those students whose personality, attitudes or reasons for studying medicine are less suitable for the medical profession.

At Linköping University 40 students are admitted to the medical programme each term, half of those are selected by local rules. In stage one all students who have Linköping as their first choice and have accepted to take part in the local admittance procedure are ranked according to average marks and SweSAT results. A number corresponding to six times the final number of admitted are invited to Linköping to write their autobiography, motivate why they want to study medicine and write a short essay on a given subject. After evaluation of the outcomes 50 per cent of these applicants are interviewed by two persons - one teacher/doctor and one layman with experience in interviewing people. After the interviews the interviewers make a common ranking of the applicants and the upper third is accepted.

At Uppsala University 55 students are admitted each term, ten of which are selected after special tests and an interview.

At Gothenburg University 57 students are admitted each term, so far, all in the central selection procedure. From next autumn, however, the selection will be locally made in Gothenburg and in a two stage procedure similar to the procedure used in Umeå.

At Lund University 82 students are admitted each term all in the central procedure. In Lund they will not start with local selection until autumn 1997.

At Karolinska Institute in Stockholm 120 students are admitted to the medical programme each term and the institute got special permission as early as 1992 to try out local admission to some of their study places. The main reason for Karolinska Institute to try out new methods for selection to their medical programme was that they felt that the selection procedures used for central admission, i.e. mark averages and SweSAT scores, when used alone, failed to give satisfactory information about the applicants' suitability for the medical profession or their motivation for medical studies.

At Karolinska Institute one third of the total number of admitted or 40 study places are allocated locally. The selection is made in three stages:

Applicants with a result of at least 1.6 on the SweSAT are invited to the Institute where they are asked to write a short essay on one of three suggested topics, a short autobiography and a motivation for their wishing to become a doctor. The chosen applicants are then interviewed twice, first by a teacher/doctor at the Institute and then by a psychologist. These interviews are semistructured and aim at finding out the applicant's motivation, maturity, judgement and intellectual mobility. The results of the interviews are evaluated and the applicants who are regarded as best suited for the the studies and the profession are chosen.

This admittance procedure was evaluated after a trial period where (1) study intermissions and drop out rates, (2) number of courses passed during the four first semesters and (3) results on the preclinical examination at the end of the fourth semester in the medical programme, were investigated. Results obtained by the locally admitted students were compared with those obtained by students who entered the programme as a result of central admission.

Students who had been locally admitted - in spite of lower average marks and test scores - performed as well as centrally admitted students. The lower limit set on the SweSAT seems to guarantee that the students possess the intellectual capacity necessary to meet the requirements of the theoretical parts of the programme.

Altogether, 415 students are selected for the medical study programmes each term. 140 of these students are admitted after some special local selection procedure usually containing two or three steps, where the SweSAT always constitute the first step and where the last step is an interview.

References

Gustafsson, J-E., Wedman, I. & Westerlund, A. (1992). The Dimensionality of the Swedish Scholastic Aptitude Test. Scandinavian Journal of Educational Research, Vol. 36, No 1.

Hindbeck, H. Hagenfeldt, K. & Åberg, H. (1994). Lokal antagning till läkarutbildning vid Karolinska institutet. Stencil, Karolinska institutet.

Holmberg, C. (1992). Antagningen till Hälsouniversitetets läkar- och sjukgymnastlinjer. Universitetet i Linköping, LiU-PEK-R-157.

Holmberg, C. (1995). Alternativ antagning till högskolan. Stencil, Linköping universitet.

Stage, C. (1992). Gender Differences on Two Instruments Used for Admission to Higher Education. To be published in Admission to Higher Education: Issues and Practice. Selected papers from the 18th IAEA Annual Conference.

Stage, C. (1993). Gender Differences on the SweSAT. A Review of Studies since 1975. Department of Educational Measurement, Umeå University, EM No 7.

Stage, C. (1993). Average School Marks and Results on the SweSAT. Department of Educational Measurement, Umeå University, EM No 4.

Stage, C. (1994). Use of Assessment Outcomes in Selecting Candidates for Secondary and Tertiary Education: A Comparison. Paper presented at the 20th Annual IAEA Conference, Wellington, New Zealand.

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

Wedman, I. (1994). The Swedish Scholastic Aptitude Test: Development, Use and Research. Educational Measurement: Issues and Practice. Vol. 13, No 2.

Translating, equating and validating Scholastic Aptitude Tests: The Israeli Case

Michal Beller

The Open University of Israel

Introduction

The ultimate goal of translating psychological tests into multiple languages is to permit cross-cultural comparisons of psychological traits and constructs among members of different cultures. At an earlier stage researchers believed they could use culture-free measures such as figural reasoning tests in assessment (Cattell, 1940) but long years of experience have taught them that there is no such thing as a culture free test or task (Frijda and Jahoda, 1966; Poortinga and Van de Vijver, 1991). Rather, there is a continuum extending from the most to the least "culturally specific" tests (Jensen, 1980). "Presumably our existing standard mental tests can be ordered along this hypothetical continuum, of course with none of them anywhere near approaching either extreme" (p. 635). Jensen suggested thinking of the degree of "culture reducedness" of a test in terms of the "cultural distance" over which a test maintains substantially the same psychometric properties of reliability, validity, item-total score correlation, and rank order of item difficulties. Since cultural distance is multidimensional, the properties of a particular test may not span the given cultural distance at all levels. A verbal test may span the cultural distance in terms of language, if accurately translated, but not the cultural distance at the conceptual level (due to different connotations in a different cultural context).

The problem of cross-cultural testing depends on whether the purpose of testing involves predictive validity or construct validity. Demonstrating useful cross-cultural validity for a particular educational or occupational criterion is invariably much easier than establishing a test's construct validity across widely differing cultures (Jensen, 1980). Whereas elimination of the verbal parts of tests tend to widen their cultural distance, it usually lowers their predictive validity when the criterion involves verbal ability, such as scholastic performance. In such cases, cross-cultural tests are more effective if they include verbal items that are appropriately translated.

This paper describes aspects of the Israeli experience regarding test translation, adaptation and calibration. In particular, this paper deals with attempts that have been made by Israel's National Institute for Testing and Evaluation (NITE), to address the issue of selecting, in a fair and valid manner, applicants to universities in Israel who are not in full command of the Hebrew language

(which is the language of instruction in all Israeli universities). The purpose of translating admissions tests is to enable meaningful comparisons, to the extent that this is possible, among applicants from different cultural backgrounds who speak different languages, regarding their prospective success in academic studies within a specific cultural milieu - that is, in Israel. The focus of the present study is not on cross- cultural comparisons or national differences. Rather, the aim is to rank-order all applicants, regardless of their mother-tongue, on a common scale, based on the Psychometric Entrance Test, that is correlated, as highly as possible, with academic success.

Casagrande (1954) presented four types of translation, differentiated according to their goals: (a) Pragmatic translation, where the sole interest lies in communicating accurately in the target language what was contained in the source language; (b) Aesthetic-poetic translation, the purpose of which is the evocation of moods, feelings, and affect in the target language that are identical to those evoked in the source language; (c) Ethnographic translation which is aimed at maintaining the meaning and the cultural content of the source language in the target language; (d) Linguistic translation which is concerned with equivalence of meanings of both morphemes and grammatical forms of the two languages.

Hulin, Drasgow and Parsons (1983) were concerned with evaluating translations of psychological instruments - ability tests; measures of attitudes, interests, etc.- that were designed to assess individual differences. They claimed that translations carried out in this area would most likely be classified as ethnographic translations, although the fit with this category is not perfect. Translators producing these translations must be familiar with both the source and target cultures as well as with the source and target languages. They must know how words and phrases are interpreted in a culture and use them appropriately in the translated version. Hulin et al's contentions seem most appropriate with respect to translating the Psychometric Entrance Test.

Description of the Psychometric Entrance Test

The Psychometric Entrance Test (PET) is a scholastic aptitude test, constructed and administered by NITE. It is used in the procedure of admissions to all Israeli universities in conjunction with a matriculation certificate, which is based on both school assessment and external nationwide achievement tests. For students of foreign origin, the school-based component is either missing or, more often, cannot be compared to the Israeli matriculation scores. Therefore, these candidates are rank-ordered on the basis of their PET score alone.

PET measures various cognitive and scholastic abilities, in an effort to estimate future success in academic studies. Similarly to SAT, PET is intended to "...measure aspects of developed ability...it makes use of the kind of basic ver-

bal and mathematical skills that develop over the years, both in and out of school. The content of the test does not reflect specific curriculums, although it is designed to be consistent with school-based learning" (Donlon, 1984, p. 58).

The test battery is comprised of three multiple-choice subtests:

1. Verbal Reasoning (V) - 60 items focusing on the verbal skills and abilities needed for academic studies: the ability to analyze and understand complex written material, the ability to think systematically and logically, and the ability to perceive fine distinctions in meaning among words and concepts. The verbal sections generally contain a number of different types of questions, such as antonyms, analogies, sentence completions, logic and reading comprehension.
2. Quantitative Reasoning (Q) - 50 items focusing on the ability to use numbers and mathematical concepts (algebraic and geometrical), to solve quantitative problems, and to analyze information presented in the form of graphs, tables and charts. Solving problems in this area requires only basic knowledge of mathematics - the math level acquired in the 9th or 10th grades in most high schools in Israel. Formulae and explanations of mathematical terms which may be needed in the course of the exam appear in the test booklet.
3. English as a Foreign Language (E) - 54 items designed to test command of the English language (reading and understanding texts at an academic level). The English subtest contains three types of questions: sentence completions, restatements, and reading comprehension. This subtest serves a dual purpose: it is a component of the PET total score, and is also used for placement of students in remedial English classes.

No correction for guessing is used in scoring the test, and examinees are encouraged to guess when they do not know the correct answer. For a more detailed description of PET and the admissions procedure to the universities in Israel, see Beller (in press).

Translated versions of PET

The variety of different native languages spoken by applicants to Israeli universities is a result of Israel's foremost national characteristic - its status as the destination of immigrants from all over the world, including, in recent years, a large number of Russian immigrants. In addition, Israel has a large Arabic-speaking minority (15% of the population). In establishing admissions policy for the universities in Israel, policy- makers and psychometricians have been faced with the problem of finding the best method for predicting the academic success of non-Hebrew-speaking applicants (along with the Hebrew-speakers) in the institutions of higher education, where the language of instruction is

Hebrew. It was decided to administer the general scholastic aptitude test in the language with which the applicant is most familiar, because it was believed that this would provide all applicants with the opportunity to demonstrate optimal performance. Therefore, PET is translated into the languages spoken by the majority of non-Hebrew-speaking applicants.

Currently, the test is translated into Arabic, Russian, English, French and Spanish. A combined Hebrew and English (H&E) version is offered to applicants who are not proficient in any of the aforementioned languages. Of the total number of examinees (56,883 in 1991/2) around 20% chose to take PET in a foreign language (10% - Arabic; 7.5% - Russian, and 2.5% - other foreign languages). The examinees who choose to take PET in a foreign language are required to take an additional Hebrew proficiency test (scored separately).

The non-Hebrew versions of PET are essentially translations of the Hebrew form, and thus have a similar structure. The English subtest is identical in all versions. The Quantitative subtest is translated and reviewed by bilingual experts. The Verbal subtest is only partially translated from the Hebrew. Most items are selected from the pool of Hebrew items, but others are specially constructed for the various language groups. For reasons of test validity an effort is made to preserve the original meaning of the test directions and, and as much as possible, of the items.

Equivalence of test items in the source and target languages means that scores derived from each of the groups taking them are comparable. In order to establish translation equivalence, both judgmental and statistical methods may be used. In the case of PET, the accuracy of the translation is checked in various ways, including translating the non-Hebrew versions back into Hebrew and comparing this back-translation with the original. Back- translation is the best known and most popular of the judgmental methods (Hambleton, 1993). Ideally, this method involves three steps (Hulin et al., 1983). The original version of the test is first translated into the target language. The target language text is then translated back into the source language by independent translators. Finally, the back-translated text is compared to the original by individuals who have not been involved in any of the previous steps. For PET this task is performed by bilingual experts who have not seen the original Hebrew text. In addition, once the test has been administered, items that do not meet specified psychometric standards are removed, post-hoc, from the test.

An essential component of culture fair testing is to ensure that all persons fully understand the requirements of each type of task involved in the test. In order to familiarize the examinees with the test, NITE publishes an information booklet which includes previously administered tests as well as explanations. This booklet is also translated into the above-mentioned languages. This procedure is particularly important, because the various language groups differ in terms of their previous experience with multiple-choice tests.

A study conducted by Gafni and Melamed (1990) indicated that the tendency to avoid guessing was found to be a function of two variables: 1) previous experience with multiple-choice tests on the part of the various language groups, and 2) the degree of familiarity of the general public with this kind of testing (assuming that such familiarity increased with the passage of time, from the time at which PET was first administered until the fourth year of operational testing). In spite of being encouraged to guess when they do not know the correct answer, only 75% to 93% of the examinees (depending on the specific subtest) responded to all the items on the test. It was postulated that different language groups might manifest different guessing behaviors. For example, it was expected that the English-speaking group would be more familiar with multiple-choice tests and, therefore, would be more likely to closely follow the test instructions. On the other hand, the Russian-speaking group, being less acquainted with this type of test, might be less inclined to guess.

The tendency to avoid guessing was measured by the proportion of two indices - two types of unanswered items: number of unreached items and omitted items. Three of five subtests (taken from an earlier version of PET) were included in the analysis: Figural Reasoning, Quantitative Reasoning and English. For each of the six dependent variables (two indices x three subtests) a covariance analysis was performed, with language group, gender, and exam date (either 1984 - the first year PET was administered, or 1987) as independent variables, and with the formula score [(Number Right)-(Number Wrong)/(k-l)] as a covariate. This score was preferred over the number right score because it was hoped that it would moderate the confounding of number-right score with the proportion of unanswered items.

A language-group effect was found for both types of unanswered items, especially, for the proportion of unreached items. Russian-, Arabic- and French-speaking examinees tended to omit more items than Hebrew-, English- and Spanish-speaking examinees in 1984; in 1987 (after four years of administering PET). Russian-speaking examinees tended to omit more items than all other groups. More unreached items were observed for the Russian-speaking group than for the other groups, both in 1984 and in 1987.

The proportions of both types of unanswered items dropped significantly from 1984 to 1987. These results were attributed to an intensive educational program being implemented among the potential examinees with respect to test preparation. An interaction effect was found for exam date with language group. While the Arabic-speaking examinees tended to omit and not to reach more items than the Hebrew-speaking examinees in 1984, they tended to answer more items (guess more) than their Hebrew-speaking counterparts in 1987.

The results suggest that people with differing cultural backgrounds differ in their tendency to guess. It is probable that some of the lower scores of certain groups on multiple-choice tests can be partially explained by these groups' tendency to avoid guessing; some of the differences in performance among the language groups can also be explained in this way. It was recommended that the importance of test instructions be emphasized, in particular among members of groups known to avoid guessing.

Equating the language versions

Translation of a test from one language to another is risky and should be done in connection with proper psychometric equating methods (Jensen, 1980; Angoff and Modu, 1973; Angoff and Cook, 1988). Words and concepts do not always take on equivalent meanings, familiarity, connotation, or difficulty level when translated into the language of another culture. The cross-cultural equating of vocabulary and other verbal translated items is accomplished by retaining only those items that maintain the same rank order of difficulty, and have the same item X total score correlations in both cultures. The purpose of this equating procedure is merely to provide comparable predictive validity for both language groups rather than to make absolute comparisons of the groups in the construct measured by the tests.

An attempt to equate scores of a test given in two languages was made by Angoff and Modu (1973) and Angoff and Cook (1988), who tried to establish score equivalencies between the verbal and mathematical scores on the College Board Spanish-language Prueba de Aptitud Academica (PAA), and the verbal and mathematical scores, respectively, of the English SAT. A set of "common" items was used as an anchor-test to calibrate and adjust for any differences between the groups in the process of equating the two tests. The data resulting from the common items were used to calibrate for differences in abilities between the two candidate groups. The two tests were then equated both by linear (Tucker or Levin) and curvilinear (equipercentile and item response theory - IRT) equating methods (see Lord, 1980 for the IRT equating method, and Angoff, 1984, for the other methods mentioned).

The basic assumption underlying these studies was that the difference between the means of the difficulty values for the two groups was a reflection of the difference in their ability levels. The researchers assumed that, basically, the test measured the same trait for both groups, and their efforts were directed at detecting those items which did not conform to the general pattern. Moreover, underlying the equating methods is the assumption that the relationship between the "common" items and the whole test is the same for the two groups.

The procedures which are used for equating the different language versions of PET to the Hebrew version are similar to the methods described above. These procedures are:

English (E) - this subtest is given to all examinees in the same language and format; therefore there is no need for calibration and the same parameters are applied in scoring the E subtest for all language versions.

Quantitative Reasoning (Q) - the general assumption for this subtest is that Math items can, in general, be translated, in a manner that makes them directly comparable. This assumption is partially checked by applying delta plot techniques (see description below as well as Angoff and Modu, 1973). The very few items which deviate extensively from the general trend of the plot are not included in the scoring procedure.

Verbal Reasoning (V) - this is clearly the most problematic area, because the meaning of verbal items may be drastically altered by translation, and therefore may not be comparable to their Hebrew counterparts. A similar equating procedure to the one described by Angoff and Modu (1973) is applied. An anchor is established by selecting items that are similar in their conventional psychometric indices and in their rank-order position among other items (using delta plot techniques) for each two groups of examinees (Hebrew and each of the foreign languages). Once an anchor is established, linear equating methods (Tucker or Levin) are applied.

Equating the different language versions is still an open question. Further research must be conducted to reveal whether the above-mentioned solution is satisfactory, or whether other equating procedures should be adopted. There is also concern that some of the groups differ greatly in average ability, so that it is unlikely that any set of common items, however appropriate, can make adequate adjustments for the differences between groups (Angoff & Cook, 1988).

The quality of the translation

In addition to proofreading, back-translation, checking for clarity of the sentences and the level of wording, the quality of the translation is assessed by using the following quantitative criteria: item analyses and item bias, reliability, validity, and test-bias.

Recently the focus of attention has recently been drawn to the Russian version, due to the wave of immigration from the former Soviet Union (at the beginning of the 1990's) that has drastically increased the number of applicants tested in Russian (e.g., 4539 in 1991 compared with 189 in 1989).

a. Item analysis and item bias

The quality of each item (in terms of its difficulty level and discrimination power) and checking for differential item functioning (DIF) of each translated item compared with the Hebrew version. The expression "differential item functioning" (or what sometimes is referred to as "item-bias") is used when referring to the simple observation that an item displays different statistical properties in different group settings (after controlling for differences in the abilities of the groups). In 1972, Angoff proposed a method for studying cultural differences, known as the delta-plot or transformed item-difficulty (TID) method. The delta-plot method calls for the calculation of item p-values (proportion correct) for each of the two groups under consideration and for the conversion of each p-value to a normal deviate, usually expressed on a scale with a mean of 13 and a standard deviation of 4. The pairs of normal deviates, one pair for each item, are then plotted on a bivariate graph with the two groups represented on the axes, each pair represented by a point. When the groups are of the same type and of the same level of ability, the plot of these points will ordinarily appear in the form of an ellipse extending from lower left to upper right, often representing a correlation of 0.98 or even higher, indicating that the rank order of difficulty of the items is essentially the same in the two groups. When the groups differ only in level of ability, the ellipse will be displaced vertically or horizontally, depending on which group has the greater ability. However, when the groups are drawn from different types of populations, the points will be dispersed in the off-diagonal direction and the correlation represented by the points will be lower. The items which fall at some distance from the plot of points, as measured by the distance of the item's bivariate point from the principal axis of the plot, may be regarded as contributing to the item X group interaction. These are the items that are clearly more difficult for one group than for the other, relative to the other items, and are ordinarily taken to be characterized by DIF (Angoff, 1993).

A study was designed by Gafni and Cnaan (1993) to detect DIF in three Russian forms of PET (the number of Russian examinees in the three versions were: 1213, 2921, 842). Table 1 presents averages and standard deviations of the item difficulty levels and discrimination indices (biserial correlations between the item score and the total score), for the Hebrew and Russian language groups.

Table 1: Difficulty levels and discrimination indices for the Hebrew and Russia language groups

		DELTA				BISERIAL				
		Mean		SD		Mean		SD		
Form	**n**	**R**	**H**	**R**	**H**	**R**	**H**	**R**	**H**	**R**
1										
V	48	12.33	11.93	2.24	2.02	.39	.41	.13	.11	.83
Q	44	11.16	11.15	2.21	1.91	.57	.67	.10	.13	.92
2										
V	47	12.37	11.71	1.98	1.91	.41	.42	.12	.09	.70
Q	44	11.74	11.25	2.09	2.15	.54	.55	.10	.09	.96
3										
V	50	11.67	10.87	2.16	2.04	.46	.44	.10	.10	.81
Q	44	10.52	11.87	2.10	1.66	.55	.53	.15	.12	.92

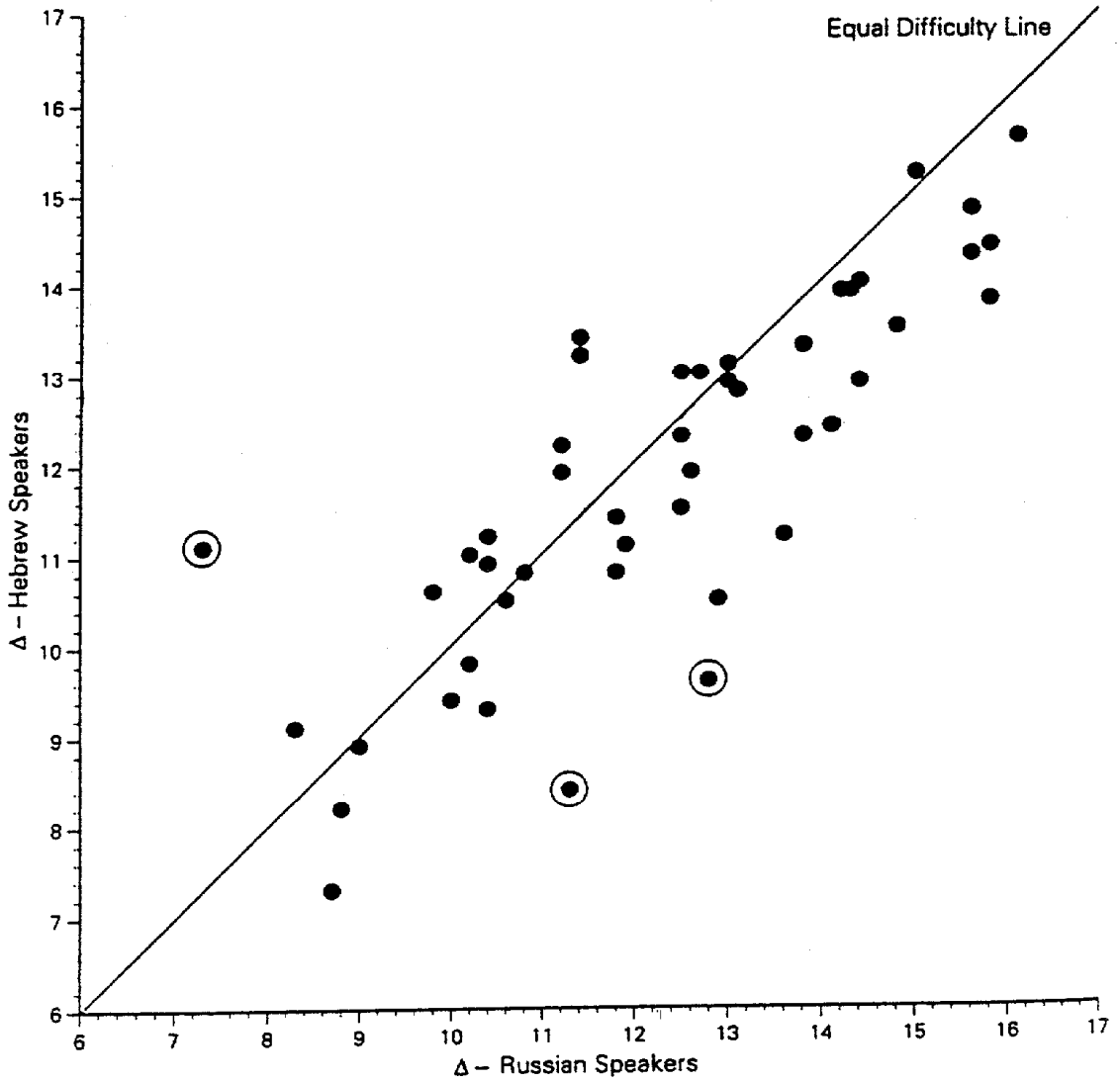
n = Number of items
R= Russian-speaking group
H=Hebrew-speaking group
r=Pearson correlation between the delta values for the H and Rgroups

The level and pattern of performance of the Russian examinees was generally very similar to that observed in the three Hebrew versions of the same test (based on 4475, 5308, and 7722 examinees respectively). The biserial correlations of the translated Russian items were found to be similar to those of the respective original Hebrew items. In addition, the similarity of the difficulty levels (as measured by deltas) was fairly high for the quantitative domain (correlations of 0.92, 0.96, 0.92 were obtained for the three Hebrew and Russian test versions). No quantitative items were detected as functioning differentially among the two groups. As can be expected, the similarity among the difficulty levels of the verbal items was lower (correlations of 0.83, 0.70, 0.81 were obtained for three Hebrew and Russian test versions). In particular, the lowest DIF value was found for the logical verbal items, while the analogies produced the largest DIF values (suggesting that the meaning of the analogy in Hebrew was not fully comparable with its meaning in Russian). The correlations between the delta values of the verbal items after deleting a few items (2 or 3) with large DIFs were greater than 0.85. These findings are consistent with Thorndike's (1973-1974), who reported a study involving the simple translation of a thirty-item reading comprehension test from English into seven other languages. The test was given to age-matched school children in eleven different countries, in the countries' national language. The single-item difficulties of the thirty items were intercorrelated among the eleven countries. The 55 resulting correlation coefficients ranged from 0.80 to 0.98, with a mean of 0.88. The average correlation, excluding pairs of countries in which the

same language was spoken, was 0.86. It can be seen that even the subtleties of a reading comprehension test survived translation quite well and that the items maintained highly similar relative difficulties across different language and national groups.

Figure 1 demonstrates the delta-plot of the verbal items on one of the above three mentioned forms of PET, for the Hebrew and Russian groups.

Figure 1: Plot of the delta values (Δ) of the verbal items included in one of the forms of PET for the Hebrew and Russian groups



Three items (all of them turned out to be analogy items) appear to show large DIFs in this delta-plot: two in favor of the Hebrew-speakers and one in favor

of the Russian- speakers. An analogy which was found to be relatively easier for the Hebrew-speakers was:

telephone book : telephone number

(1) phonograph record : sound

(2) dictionary : definition

(3) atlas : city

(4) encyclopedia : knowledge

A probable explanation for this finding is that typical Russian dictionaries contain words, but not definitions. They are used for translation from Russian to other languages and vice versa, but not as Russian-Russian dictionaries. These differences led many Russian examinees to choose distractor (3) as the correct answer.

Another example of a relatively difficult analogy for the Russians-speakers was:

thermometer : medication

(1) pressure gauge : pressure

(2) speedometer : brakes

(3) weighing scale : malnutrition

(4) compass : north

The word "speedometer" (which has a straightforward meaning in Hebrew and in English) is used as a Latin word in Russian, and therefore it is more difficult. It was hypothesized that if that was the case, then Russian men would perform relatively better on this analogy than Russian women. Indeed, the difference in performance between men and women on this item was 1.5% in Hebrew and 9% in Russian.

The following analogy was found to be relatively easier for the Russian-speakers:

plough : furrows

(1) chalk : lines

(2) brush : dirt

(3) oar : water

(4) car : road

It turns out that the word "furrows" occurs more frequently in Russian than in Hebrew, and this may explain the direction and magnitude of the DIF value that was obtained.

b. Reliability

The internal reliability of each subtest, as well as that of the total score, was estimated. Table 2 presents the median internal consistency coefficients (KR-20) for the three subtests and the total score, for the various language versions of PET (as mentioned above these are in: Hebrew, Arabic, Russian, English/Heb, Spanish, and French).

Table 2: Median reliability coefficients (KR-20) of PET subtests and of the composite total score for each language version

	V	**Q**	**E**	**PET**
Hebrew(25)	0.89	0.90	0.93	0.95
Russian (7)	0.86	0.88	0.90	0.94
Arabic(5)	0.68	0.86	0.82	0.91
Hebrew/	0.89	0.89	0.95	0.95
Spanish(2)	0.77	0.87	0.92	0.92
France(2)	0.78	0.87	0.88	0.91

The number of existing versions are in parenthesis.

These reliabilities are relatively high, both for the Hebrew and for the other language versions. The somewhat lower reliability of the Verbal Reasoning subtest (especially within the foreign languages) may be partially explained by the heterogeneity of this subtest and partially by problems in the fidelity of the translation.

The lowest reliability was found for the Arabic version, and this may be related to differences in ability level. Internal reliability is not solely determined by the quality of the test items and the quality of the translation, but also by the true variance within the group of examinees. A test which is too easy or too difficult for a particular subgroup would be less reliable. From experience accumulated at NITE, it seems that, in many cases, the quality of the translation is confounded with differences in ability level. When two groups differ in ability level, this in and of itself creates differences in reliability, comparability and item-DIF. When items are too difficult for a certain group the reliability of the test for that group is relatively low. In light of this, a Verbal Reasoning test was specially constructed for the Arabic version by including much easier items. While this new subtest had a higher reliability, in adaptation, it probably introduced a larger error of equating than that which existed in the old subtest.

c. Validity

The predictive validity of the selection procedure is routinely tested against the criterion of success (GPA) at the end of first year university studies and at the end of undergraduate studies. The validities of PET's total score (corrected for range restriction) are 0.53 for Liberal Arts, 0.50 for Science, 0.45 for Social Sciences and 0.43 for Engineering, with an average validity of 0.46 across all areas of study (Oren, 1992).

Validity studies (both construct and predictive) are being carried out for the translated versions (provided that a large enough sample exists). In a recent study (Kennet-Cohen, 1993), the validity of the PET score was calculated for the Russian-speaking group (N=772) and compared to that of the Hebrew speakers (N=2410). Across all fields of study, the average validity coefficients of PET within the Russian group (calculated for translated Russian versions of PET) were found to be similar to those of the Hebrew group. Within fields of study, the validity of PET for the Russian group was found to be relatively lower than that of the Hebrew group in the Humanities, Social Sciences and Nursing, but relatively higher in the Exact Sciences, Natural Sciences and Engineering.

d. Test-bias

The question of test-bias was studied in Israel for different language groups as well as for other groups. The term "bias" has, in the psychometric literature, a narrow technical definition. It refers to systematic errors in the predictive validity or construct validity of test scores of individuals that are associated with the individual's group membership. The assessment of bias is a purely objective, empirical, statistical and quantitative matter, entirely independent of subjective value judgments and ethical issues (Jensen, 1980). According to Cleary (1968), a test is defined as biased against a group if it consistently under-predicts criterion scores for members of that group. In general, no substantial under-prediction of criterion scores of members of minority groups was found, although the groups differed on the predictor, as well as on the criterion, scores (see, Baron and Gafni, 1988; Beller and Ben- Shakhar, 1983; Kennet, Oren and Pavlov, 1988; Zeidner, 1986, 1987).

Recently, research efforts have been made at NITE to determine whether test bias exists for the Russian-speaking group of examinees. Results from research carried out by Kennet-Cohen (1993) demonstrate that PET tends to over-predict Russian-speakers' GPA in the faculties of Humanities, Social Sciences, and Nursing. In Engineering no prediction bias was found, and in the Natural Sciences a slight under-prediction of the Russian-speakers' GPA was detected. It was hypothesized that over-prediction of Russian-speakers' GPA is observed in fields of study which are verbally loaded, and require a better mastery of

Hebrew. Therefore, it may be expected that this over-prediction will gradually disappear in the coming years, after proficiency in Hebrew is attained by this group.

Unlike the use of predictive validity to evaluate the quality of the equating method, validity can be integrated into the equating design itself, as suggested by Wainer (personal communication). He suggested equating different language versions by "anchoring" them via a common criterion score. This solution is not essentially different from the examination of fairness of prediction for different groups. Although this kind of investigation is fundamental to the process of professional test construction and validation, the idea of using group variables in scoring seems ethically unacceptable.

It is claimed that such a procedure would be publicly indefensible; moreover, it would be harmful to the goal of achieving a fair and just selection system. Furthermore, group membership should not be used as a predictor because it is only indirectly related to the criterion. Certain individuals may not perform well on the criterion, not because they belong to a certain group, but because they are low on some trait that happens to correlate with group membership. Efforts should be focused at directly measuring those traits. A crucial difference between group membership and any ability measure is that an individual can never change his or her group membership, whereas ability and knowledge can be improved with effort.

Summary

PET is translated from Hebrew into the five languages (Arabic, Russian, English, French and Spanish) spoken by the majority of non-Hebrew-speaking applicants to Israeli universities. This is rather a unique endeavor, which demands a major professional and financial investment.

From a psychological viewpoint, the task of making cross-language comparisons of the kind needed for admissions decisions is highly complex. One may argue that this task is essentially impossible, particularly when differences in ability between the various language groups are large. It cannot be automatically assumed that the translated items will have the same meaning and relative difficulty for the various language groups as they had on the original Hebrew version. This assumption needs to be carefully checked.

An attempt is currently being made to equate the different language versions to the Hebrew versions, so that all examinees may be rank-ordered on the same scale, regardless of which language version they took. The issue of equating different language versions clearly requires further research which may reveal whether the equating procedure which has been adopted is satisfactory, or whether different equating procedures should be used. However, regardless of

what specific equating method should be adopted, it is the conviction of the authors that administering the test in the examinee's native language, and then applying even a sub-optimal equating technique, is far more appropriate than the alternative of administering the Hebrew version to all language groups.

The research gathered so far by NITE suggests that investment of time, effort and money in translating and adapting admissions tests may produce satisfactory results, in terms of reliability, validity and test-bias.

References

Beller, M. (1995). Translated versions of Israel's inter-university Psychometric Entrance Test (PET). In T. Oakland, and R. K. Hambleton (Eds.). International Perspectives on Academic Assessment, 207-218. Boston: Kluwer.

Beller, M., and Gafni, N. (1995). Equating and validating translated scholastic aptitude tests: The Israeli case. In G. Ben-Shakhar, and A. Lieblich (Eds.). Studies in Psychology: A volume in honor of Sonny Kugelmass. Scripta Hierosolymitana, 36, 202-219. Jerusalem: Magnes Press.

Angoff, W. H. (1984). Scales norms and equivalent scores. Princeton, NJ: Educational Testing Service. Reprint of chapter in Educational Measurement, 2d., ed. R.L. Thorndike. Washington, D. C.: American Council on Education, 1971.

Angoff, W. (1993). Perspectives on differential item functioning methodology. In Holland, P. W., & Wainer, H. (Eds.). Differential item Functioning. Lawrence Erlbaum Associates, Hillsdale, NJ. 3-24. Angoff, W. H., & Modu, C. C. (1973). Equating the scales of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. Research Report 3. New York: College Entrance Examination Board.

Angoff, W. H., & Cook, L. L. (1988). Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test. College Board Report, No. 88-2, New- York, New-York.

Angoff, W. H., & Ford, S. F. (1973). Item-rate interaction on a test of scholastic ability. Journal of Educational Measurement, 10, 95-106.

Baron, H., & Gafni, N. (1989). An examination of item and criterion- related bias for Hebrew and Arabic speaking examinees in Israel. N.I.T.E., Report #93, Jerusalem, Israel. A paper presented in the AERA Conference, San Francisco, 1989.

Beller, M. (1994). Psychometric and social issues in admissions to Israeli universities. Educational Measurement: Issues and Practice, 13(2), 12-20

Beller, M., & Ben-Shakhar, G. (1983). On the fair use of psychological tests. Megamot, 28,42-56. (in Hebrew).

Casagrande, J. (1954). The ends of translation. International Journal of American Linguistics. 20, 335-340.

Cattell, R. B. (1940). A culture-free intelligence test: Part I. Journal of Educational Psychology, 31, 161-179.

Cleary, T. A. (1968) Test bias: prediction of grades of negro and white students in integrated colleges. Journal of Educational Measurement, 5,115-124.

Donlon, F. T. (Ed.) (1984). The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests. New-York: College Entrance Examinations Board.

Frijda, N., & Jahoda, G. (1966). On the scope and methods of cross cultural research. International Journal of Psychology, 1, 109-127.

Gafni, N., & Cnaan-Yehoshafat, Z. (1993). An examination of differential item functioning for Hebrew and Russian-speaking examinees in Israel. A paper presented at the annual conference of the Israeli Psychological Association, Ramat-Gan.

Gafni, N., & Melamed, E. (1990). Differential Tendencies to Guess as a Function of Gender and Lingual-cultural reference group. A paper presented at the annual conference of the American Educational Research Association, Boston, 1990.

Hambleton, R. (1993). Translating achievement tests for use in cross-national studies. European Journal of Psychological Assessment, 9, 57-68.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Applications to psychological measurement. Homewood, IL:Dow Jones Irwine.

Jensen, A. R. (1980). Bias in mental testing. London:Methuen; New-York: Free Press.

Kennet, T., Oren, C. & Pavlov, Y. (1988). Analysis of the culture fairness of the selection-procedure in two Israeli universities. N.I.T.E., Report #78, Jerusalem, Israel. (in Hebrew).

Knnet-Cohen, T. (1993). An examination of predictive bias: the Russian version of the Psychometric Entrance Test for Israeli universities. Paper presented at the International Test Commission conference, Oxford, England. N.I.T.E., Report #177, Jerusalem, Israel.

Lord, F. M. (1980). Application of Item Response Theory to Practical Testing Problems. Hillsdale, N.J.: Erlbaum.

Oren, C. (1992). On the validity of PET: a meta analyses (1984-1989). N.I.T.E., Report #160, Jerusalem, Israel. (in Hebrew).

Poortinga, Y., H., & Van De Vijver, F. J. R. (1991). Culture-free measurement in the history of cross-cultural psychology. Bulletin of the International Test Commission, 18, 72-87.

Thorndike, R. L. (1973-1974). Reading as reasoning. Reading Research Quarterly, 9, 135-147.

Zeidner, M. (1986). Sex differences in scholastic ability in Jewish and Arab college students in Israel. Journal of Social Psychology, 7, 847-852.

Zeidner, M. (1987). A test of the cultural bias hypothesis: Some Israeli findings. Journal of Applied Psychology, 72, 38-48.

Erfolgsprognose in medizinischen Studiengängen - Zur Validität des Tests für medizinische Studiengänge und anderer Auswahlinstrumente

Eckhard Klieme

Institut für Test- und Begabungsforschung, Bonn (Deutschland)

1. Der Test für medizinische Studiengänge (TMS) und seine Verwendung bei der Zulassung zum Studium in Deutschland

Der Test für medizinische Studiengänge (TMS) wurde Ende der 70er Jahre am Institut für Test- und Begabungsforschung entwickelt und in den Jahren 1980 bis 1985 einer umfangreichen Erprobung unterzogen. Seit 1986 sind in der Bundesrepublik Deutschland (mit wenigen Ausnahmen) alle Bewerber für die Studiengänge Medizin, Zahnmedizin und Veterinärmedizin verpflichtet, am Test teilzunehmen.

Zu jedem Testtermin wird eine neu entwickelte Version des TMS vorgelegt. Der Aufbau des Tests bleibt jedoch unverändert: Gegliedert in mehrere Untertests, sind insgesamt 184 Aufgaben nach dem Multiple-choice-Prinzip zu beantworten; hinzu kommt ein spezielles Verfahren zur Messung der Fähigkeit, konzentriert und sorgfältig zu arbeiten. Die vier wichtigsten Aufgabengruppen (sie allein füllen vier der insgesamt fünf Stunden Bearbeitungszeit) erfassen die Fähigkeit, in medizinischen und naturwissenschaftlichen Kontexten schlußfolgernd zu denken, z.B. Graphiken und Tabellen zu interpretieren, längere Textpassagen zu verstehen sowie quantitative und formale Probleme zu lösen. Einen zweiten Schwerpunkt bilden Aufgabengruppen, die verschiedene Aspekte der Verarbeitung visueller Informationen überprüfen, z.B. das räumliche Vorstellungsvermögen, die Wahrnehmungsgenauigkeit sowie Konzentration und Sorgfalt beim Erkennen visueller Zeichen. Als dritter Bereich wird die Merkfähigkeit (sowohl für sprachliche als auch für bildliche Informationen) erfaßt.

Der Test ist auf die Prüfung der *Studien*eignung ausgerichtet; seine Aufgaben sollen die zentralen Anforderungen des Studiums abbilden. Inwieweit Erfolg im Test bzw. im Studium auch mit späterem beruflichem Erfolg einhergeht, läßt sich angesichts der Vielzahl höchst unterschiedlicher medizinischer Berufsfelder und mangels eindeutiger Kriterien des „Berufserfolgs“ kaum empirisch klären.

Bei der Zulassung zu den medizinischen Studiengängen wird seit dem Wintersemester 1986/87 folgendermaßen verfahren: 45 Prozent der verfügbaren Studienplätze werden nach der Kombination von Abitur-Durchschnittsnote und Testleistung vergeben, d.h. in dieser „Abitur/Test-Quote“ kommen jene Bewerber zum Zuge, welche die besten „Wertzahlen“ (berechnet als im Ver-

hältnis 55:45 gewichtete Summe von Abitur- und Testergebnis) mitbringen. Weitere 10 Prozent der Plätze gehen an jene Bewerber, die (nach „Abarbeitung“ der Abitur/Test-Quote !) die besten TMS-Resultate vorweisen können. 20 Prozent der Studienplätze werden sodann nach Wartezeit, 15 Prozent nach dem Ergebnis von Auswahlgesprächen an den Hochschulen vergeben. Die restlichen Plätze (ca. 10 Prozent) bleiben bestimmten Gruppen von Bewerbern (Ausländern, sog. Härtefällen u.a.) vorbehalten. Wer in keiner dieser „Quoten“ berücksichtigt wurde, kann sich in späteren Semestern beliebig oft erneut bewerben.

2. Fragestellung und Datenbasis der Validitätsstudie

Das Institut für Test- und Begabungsforschung ist nicht nur mit der Entwicklung sondern auch mit der Evaluation des TMS betraut. Eine der wichtigsten Fragestellungen der Evaluation ist, wie gut jene Studierenden, die nach den oben beschriebenen verschiedenen Kriterien („Quoten“) zugelassen worden sind, das Studium bewältigen. Sind beispielsweise die Studierenden, die aufgrund guter Abitur- und Testleistungen zugelassen wurden, in den Examina erfolgreicher als jene, die ihre Studienplätze erst nach einer langen Wartezeit erhielten? Darüber hinaus sollte die hier vorzustellende Studie prüfen, inwieweit sich Leistungen in mündlichen und schriftlichen Examina überhaupt aus Schul- und Testleistungen vorhersagen lassen.

In einer umfangreichen Längsschnitt-Untersuchung wurde der Studienverlauf von insgesamt 28 000 Personen erfaßt. Es handelt sich um all jene Personen, die in den Jahren 1986 oder 1987 (also nachdem der TMS verbindlich geworden war) am Test teilgenommen haben, in einem der folgenden Jahre zum Studium der (Human-) Medizin zugelassen worden sind und sich bis zum Stichtag 31.12.1992 der Ärztlichen Vorprüfung (ÄVP) gestellt haben, die frühestens nach vier Semestern abgelegt werden kann.

Für jede dieser Personen wurden folgende Daten erfaßt: die Durchschnittsnote im Abitur, die Leistungen im TMS (Gesamttest sowie Untertests), das Kriterium der Zulassung, die Leistungen in der ÄVP (mündliche Prüfung sowie Fächer der schriftlichen Prüfung, die aus Multiple-choice-Fragen besteht), ggfs. auch aus Wiederholungsprüfungen, sowie die Studiendauer bis zum Examen.

3. Studienerfolg in Abhängigkeit vom Auswahlkriterium

Als „erfolgreich“ soll hier gelten, wer die Ärztliche Vorprüfung nach der Mindest-Studienzeit, d.h. nach vier Fachsemestern, im ersten Anlauf besteht. Von den 28 000 untersuchten Studentinnen und Studenten haben 66 Prozent die ÄVP als wichtigste „Hürde“ im Studium in diesem Sinne erfolgreich gemei-

stert. Unter jenen, die in der Abitur/Test-Quote zugelassen worden sind, betrug die Erfolgsrate sogar 80 Prozent, in der „Test-Quote“ 62 Prozent, bei Zulassung nach Wartezeit 45 Prozent und unter denjenigen, die ihre Zulassung dem Ergebnis eines Auswahlgesprächs verdanken, 49 Prozent.

Die Unterschiede zwischen den Gruppen, die nach verschiedenen Kriterien zugelassen wurden, sind sehr bedeutsam. Zu beachten ist jedoch, daß diese Kriterien nicht voneinander unabhängig sind. Beispielsweise werden Auswahlgespräche nur mit Kandidaten geführt, deren Abitur- und Testleistungen nicht für eine Zulassung ausreichen. Angesichts dieser „Vorselektion“ spricht es für die Validität des Interviews, daß es noch höhere Erfolgsraten ergibt als die Auswahl nach Wartezeit.

Nur in Simulationsrechnungen lassen sich „absolute“ Vergleichsmaßstäbe entwickeln. So konnten wir beispielsweise abschätzen, daß bei einer Zufalls-Auswahl, also bei einer Vergabe der Studienplätze mittels Losentscheids, ganze 48 Prozent der Zugelassenen im obern erwähnten Sinne erfolgreich wären. Das praktizierte Verfahren - insbesondere die Berücksichtigung von Testleistungen und Abiturnoten - erhöht demnach die Erfolgsrate um nahezu 20 Prozentpunkte. (Über weitere Simulations-Ergebnisse berichtet Günter Trost in seinem Symposiumsbeitrag.)

4. Zusammenhang zwischen Testleistung bzw. Abiturnote einerseits und Leistungen in medizinischen Prüfungen andererseits

Die Beziehung zwischen dem TMS-Ergebnis einerseits und der Gesamtnote der Ärztlichen Vorprüfung andererseits läßt sich recht gut durch eine lineare Gleichung beschreiben. Die Enge dieses Zusammenhangs beziffert der Validitätskoeffizient, der als Korrelationsmaß Werte zwischen 0 und 1 annehmen kann. In unserer Studie ergab sich ein Koeffizient von 0,45. Im Vergleich zu internationalen Forschungsresultaten, die für Studieneignungstests durchweg Validitätskoeffizienten zwischen 0,30 und 0,60 ermitteln, kann die Validität des TMS als hoch bewertet werden, wenn man berücksichtigt, daß hier der Studienerfolg über die relativ große Zeitspanne von mindestens zwei Jahren hinweg prognostiziert wird.

Die Korrelation zwischen der Abitur-Durchschnittsnote und der ÄVP-Gesamtnote läßt sich mit 0,47 beziffern. Auch dieser Zusammenhang ist also relativ eng. Von großer praktischer Bedeutung ist nun, daß die Kombination von Testergebnis und Schulabgangsnote, die sog. Wertzahl, mit 0,54 eine deutlich höhere Validität besitzt. Offensichtlich erfassen Test und Schulnote unterschiedliche Aspekte der Studieneignung, und durch ihre Kombination kann die Qualität der Erfolgsprognose deutlich verbessert werden.

Interessant ist ferner eine getrennte Betrachtung des mündlichen und des schriftlichen Teils der ÄVP. Die mündliche Examensleistung ist weniger gut prognostizierbar als die schriftliche, wie die Koeffizienten von 0,36 respektive 0,57 für die Korrelation mit der „Wertzahl“ anzeigen. Hierfür ist nicht zuletzt die geringere Meßgenauigkeit mündlicher Prüfungen verantwortlich, und es gibt Anzeichen für einen „Methoden-Effekt“: Die Note im mündlichen Examen hängt mit der Abiturnote, in die ja auch mündliche Leistungen eingehen, etwas enger zusammen als mit dem Testergebnis; beim schriftlichen Examen ist es umgekehrt.

Frühere Untersuchungen, die in den 80er Jahrer an kleineren Stichproben durchgeführt wurden, zeigten nahezu identische Resultate. Wir konnten damals feststellen, daß auch bei nachfolgenden Prüfungen im klinischen Studienabschnitt, nach fünf bis sechs Studienjahren, noch ein bedeutsamer Zusammenhang zwischen Zulassungskriterien und Prüfungsergebnis besteht. Die Korrelationskoeffizienten liegen dann etwa um 0,10 niedriger, das Muster ist identisch.

Neueste Untersuchungen zeigen, daß der Erfolg in den Studiengängen Zahnmedizin und Veterinärmedizin etwa ebenso gut prognostizierbar ist wie im Fach Medizin.

5. Ein Strukturmodell zur Vorhersage des Erfolgs in medizinischen Studiengängen

Abschließend sollen innerhalb des Tests für medizinische Studiengänge verschiedene Teilbereiche unterschieden werden. Wir verwenden hierzu ein sogenanntes Strukturgleichungsmodell, dessen statistische Grundlagen an dieser Stelle nicht näher erläutert werden können (vgl. hierzu die Arbeitsberichte des Instituts für Test- und Begabungsforschung). Die wesentliche Idee besteht darin, die verschiedenen Untertests des TMS zu sogenannten Faktoren zusammenzufassen, die gleichsam idealtypische, von Meßfehlern befreite Leistungsdimensionen darstellen. Wie in Abschnitt 1 angedeutet, lassen sich im TMS die Faktoren (1) „Schlußfolgerndes Denken“, (2) „Visuelle Informationsverarbeitung“ und (3) „Merkfähigkeit“ unterscheiden. Die sechs Fächer im schriftlichen Teil der Ärztlichen Vorprüfung (Physik, Chemie/Biochemie, Biologie, Physiologie, Anatomie und Medizinische Psychologie/Soziologie) bilden demgegenüber einen einzigen Faktor, innerhalb dessen keine spezifischen Leistungsdimensionen unterschieden werden können.

Das Modell, das eine gute Anpassung an die vorliegenden Daten aufweist, ergab folgende Koeffizienten für die Korrelationen mit dem ÄVP-Faktor: 0,62 beim ersten TMS-Faktor, 0,23 beim zweiten und 0,32 beim dritten Faktor. Die Prognosekraft des TMS beruht demnach überwiegend auf der Erfassung von Fähigkeiten zum schlußfolgernden Denken in medizinischen und naturwissen-

schaftlichen Kontexten. Das Konzept des Tests, das sich wesentlich auf die „Simulation“ komplexer Anforderungen des Studiums (z.B. des Verstehens längerer Texte, der Interpretation von Graphiken und Tabellen, des Operierens mit Zahlen, Einheiten und Formeln) stützt, hat sich demzufolge bewährt.

Testergebnisse versus Schulnoten als Auswahlkriterien: Paternoster-Effekt, Filter-Effekt, Kosten-Nutzen-Effekte und Auswirkungen auf die Fairneß der Zulassung

Günter Trost

Institut für Test- und Begabungsforschung, Bonn (Deutschland)

Für die Beurteilung der Brauchbarkeit von Studierfähigkeitstests bei der Zulassung zu bestimmten Studiengängen ist eine Reihe von Aspekten bedeutsam. Einer der wichtigsten Aspekte ist die Prognosekraft derartiger Testverfahren bezüglich des Erfolgs in den betreffenden Studiengängen; Ergebnisse empirischer Untersuchungen zu dieser Frage finden sich in den Beiträgen von Michal Beller und Eckhard Klieme. Weitere zentrale Fragen zielen auf (a) die Beziehung zwischen Testleistungen und Schulleistungen und die Folgen, die sich aus der Enge dieser Beziehung ergeben, wenn das eine Auswahlkriterium durch das andere ersetzt wird, (b) den Einfluß, den das Ergebnis eines Studierfähigkeitstests auf die Studienwahl der Testteilnehmer ausübt, (c) das Verhältnis von Kosten und Nutzen der Entwicklung und Verwendung eines Studierfähigkeitstests und (d) die Fairneß eines Auswahlverfahrens aufgrund der Leistungen in einem Test etwa im Vergleich mit der Fairneß einer Zulassung allein aufgrund der Schulabschlußnote. Antworten auf diese Fragen, wie sie sich aus den Begleituntersuchungen zum Test für medizinische Studiengänge in der Bundesrepublik Deutschland ergeben, enthält der vorliegende Beitrag.

1. Wie eng ist die Beziehung zwischen der Leistung im Test für medizinische Studiengänge und der Schulleistung, und wie wirkt sie sich aus?

Der Kennwert für die Enge der Beziehung zwischen dem Gesamtergebnis im Test für medizinische Studiengänge (TMS) und der Durchschnittsnote im Abiturzeugnis beträgt 0,40. (Es handelt sich um einen Korrelationskoeffizienten, der auf einer Skala von 0,00 - keinerlei Zusammenhang - bis 1,00 - totaler gleichsinniger Zusammenhang - variieren kann.) Ein Kennwert in dieser Höhe bezeichnet einen mäßigen Zusammenhang. Das bedeutet: Im Test kommen überwiegend Fähigkeiten zum Ausdruck, die in der schulischen Leistungsbewertung keinen Niederschlag finden, die jedoch wichtig für den Erfolg in den

medizinischen Studiengängen sind, wie die Ergebnisse der Bewährungskontrollen belegen (siehe das Referat von Eckhard Klieme).

Verwendet man eine Kombination von Abiturdurchschnittsnote und Testergebnis als Kriterium für die Zulassung zum Medizinstudium, wie das in der Bundesrepublik Deutschland bei der „Haupt-Zulassungsquote“ geschieht, so stellt man folgendes fest: Circa 30 Prozent der zugelassenen Bewerber verdanken diese Zulassung ihrer guten Testleistung; sie hätten keinen Studienplatz erhalten, wäre die Auswahl allein aufgrund der Durchschnittsnote erfolgt („Paternoster-Effekt“).

2. Hat das Testergebnis Einfluß auf die Entscheidung der an einem Medizinstudium Interessierten, ob sie sich tatsächlich um die Zulassung zu einem medizinischen Studiengang bewerben oder nicht?

Alle Personen, welche die allgemeine Hochschulreife erworben haben oder Schüler bzw. Schülerinnen der Jahrgangsstufe 13 des Gymnasiums sind, können in der Bundesrepublik Deutschland am TMS teilnehmen. Die Teilnahme ist kostenlos. Wer sich um einen Studienplatz in den medizinischen Studiengängen bewirbt, muß zuvor am Test teilgenommen haben.

In einer Längsschnittuntersuchung über einen Zeitraum von 4 Jahren wurde das Bewerbungsverhalten aller Personen beobachtet, die im Herbst 1986 am TMS teilgenommen hatten. Dabei stellte sich heraus:

(a) Während der vier Jahre nach der Teilnahme am Test bewarben sich 34 Prozent der einstigen Testteilnehmer kein einziges Mal um die Zulassung zu einem der drei medizinischen Studiengänge.

(b) Im Mittel lag die Testleistung jener, die sich in der Folgezeit nicht bewarben, deutlich niedriger als die Testleistung jener, die sich um einen medizinischen Studienplatz bewarben. (Die Schulleistungen der späteren Bewerber waren demgegenüber im Mittel nur geringfügig besser als die Schulleistungen der späteren „Nicht-Bewerber“.)

Der Test übt mithin einen erwünschten „Filter-Effekt“ aus: Das Testergebnis hat Einfluß auf die Entscheidung der Teilnehmer, ob sie sich um die Zulassung zu einem der medizinischen Studiengänge bewerben oder nicht. Diese „Selbst-Auslese“ entlastet den institutionellen Auswahlprozeß.

3. Wie ist das Verhältnis von Kosten und Nutzen, wenn ein Studierfähigkeitstest von der Art des TMS verwendet wird?

Die Kosten eines Testverfahrens wie des TMS lassen sich leicht beziffern. In der Bundesrepublik Deutschland belaufen sie sich auf umgerechnet ca. 2 Mio. Schweizer Franken pro Jahr. Der Nutzen eines solchen Testverfahrens ist dagegen zum Teil immaterieller Natur und läßt sich deshalb nur schwer in Geldbeträgen quantifizieren.

Einer der in Betracht kommenden Nutzenaspekte läßt sich allerdings - unter den in der Bundesrepublik gegebenen Auswahl- und Studienbedingungen - in tatsächliche Kosteneinsparungen umrechnen, wie das folgende Beispiel zeigen soll. Anhand der Daten aus der Bewährungskontrolle mit etwa 28 000 Studierenden der Medizin (siehe das Referat von Eckhard Klieme) läßt sich in Modellrechnungen der Prozentsatz der Personen bestimmen, die im ersten Anlauf die Ärztliche Vorprüfung - nach wie vielen Semestern auch immer - bestünden, wenn die Studienplätze nach Zufall vergeben würden, wenn also keinerlei systematische Auswahl stattfände. Das hier angelegte Erfolgskriterium ist mithin grosszügiger definiert als bei der Modellrechnung im Beitrag von E. KLIEME. Diese „Basisrate“ beträgt 69 Prozent. Wählte man statt dessen allein aufgrund der Ergebnisse im Test für medizinische Studiengänge aus, so läge die Erfolgsrate unter den Zugelassenen bei 90 Prozent. Der Anteil derer, die mindestens ein Semester länger studieren, weil sie die Ärztliche Vorprüfung wiederholen müssen, verringert sich mithin um 21 Prozentpunkte, das sind in der Bundesrepublik Deutschland jährlich 1 650 Personen.

Die Kosten eines kompletten Medizinstudiums für den Steuerzahler werden in der Bundesrepublik auf umgerechnet ca. 320 000 Schweizer Franken beziffert (Studiengebühren werden nicht erhoben); das entspricht pro Studienplatz und Semester einem Betrag von 27 000 Schweizer Franken. Durch die Verwendung des Testergebnisses als Auswahlkriterium anstelle einer Zufallsauswahl und die dadurch bewirkte Verbesserung der Erfolgsrate können - unter den in Deutschland geltenden Bedingungen - jährlich umgerechnet knapp 45 Mio. Schweizer Franken an Studienkosten eingespart werden, denen die 2 Mio. Schweizer Franken für die Testentwicklung und Testdurchführung gegenüberstehen. Bestimmt man nur den Zuwachs in der Erfolgsrate, der eintritt, wenn man das Testergebnis zur Abiturdurchschnittsnote als Auswahlkriterium hinzunimmt, so gelangt man zu einer Verbesserung um vier Prozentpunkte; diese Differenz entspricht noch immer einer jährlichen Einsparung an Studienkosten in Höhe von 8,5 Mio. Schweizer Franken.

4. Wie fair ist ein Auswahlverfahren aufgrund der Testleistung gegenüber einem Auswahlverfahren aufgrund der Abiturdurchschnittsnote?

Die Frage der Fairneß des Verfahrens bei der Zulassung zum Medizinstudium soll im Blick auf die Gruppen der männlichen und der weiblichen Bewerber erörtert werden.

Zunächst ein paar Fakten:

Die Abiturdurchschnittsnote der weiblichen Bewerber um Plätze in den medizinischen Studiengängen ist im Mittel geringfügig besser als die Abiturnote der männlichen Bewerber (Unterschied: 0,12 Noteneinheiten auf der Skala von 1,0 - bester Wert - bis 4,3 - niedrigster Wert -; die Standardabweichung beträgt für beide Gruppen 0,6 Noteneinheiten).

Der Gesamtwert im Test für medizinische Studiengänge liegt im Mittel bei den männlichen Bewerbern etwas höher als bei den weiblichen (Unterschied: 2,2 Standardpunkte auf einer Skala, die von 70 bis zu 130 Punkten reicht; der Mittelwert über alle Testteilnehmer ist 100; die Standardabweichung beträgt für beide Gruppen 10 Standardpunkte).

Aus diesen Daten könnte man voreilig schließen, bei einer Zulassung allein aufgrund der Abiturdurchschnittsnote seien Männer, bei einer Zulassung allein aufgrund der Testleistung seien Frauen benachteiligt. Diesen Schlußfolgerungen liegt ein Verständnis von Fairneß der folgenden Art zugrunde: „Ein Auswahlverfahren ist fair, wenn der Anteil der Mitglieder definierter Teilgruppen unter den Ausgewählten gleich dem Anteil dieser Teilgruppen in den Gesamtgruppe der Bewerber ist.“ Diese Definition läßt jedoch die tatsächliche Eignung der Mitglieder der jeweiligen Teilgruppen, also ihren zu erwartenden Studienerfolg, außer acht.

In der Psychologie und den Sozialwissenschaften werden deshalb andere Definitionen von Fairneß bevorzugt, welche die Eignung der Mitglieder der betreffenden Teilgruppen berücksichtigen. Eine derartige Definition lautet: „Ein Auswahlverfahren ist fair gegenüber bestimmten Bewerbergruppen, wenn Teilgruppen, die jeweils gleiche Erfolgsaussichten haben, die gleiche Zulassungschance erhalten.“ Ein Blick auf die Studienleistungen von Frauen und Männern (die nicht nach Schulnoten oder Testergebnissen ausgelesen sind) zeigt, daß etwa in der Ärztlichen Vorprüfung Männer im Mittel einen Standardwert erzielen (ebenfalls auf der Skala von 70 bis 130 Punkten; Standardabweichung um 9 Punkte), der um knapp drei Punkte über jenem liegt, den die weiblichen Studierenden erzielen.

Diesen höheren Erfolgsaussichten der Männer wird das Auswahlverfahren nur in unzureichendem Maße gerecht. Setzt man nämlich Schul- bzw. Testleistung jeweils in Beziehung zur Studienleistung, so stellt man folgendes fest:

- Männliche Bewerber sind, gemessen an ihrer Studienleistung, in beiden Fällen benachteiligt, also sowohl, wenn allein aufgrund der Abiturdurchschnittsnote, als auch, wenn allein aufgrund des Testergebnisses zugelassen wird.
- Das Ausmaß an „Unfairneß“ ist jedoch deutlich geringer, wenn das Testergebnis als Auswahlkriterium verwendet wird.

Nutzen, Fairness, Validität und Akzeptanz von Selektionsverfahren

Urs Schallberger

Universität Zürich, Psychologisches Institut

Die Angewandte Psychologie befasst sich seit ihren Anfängen zu Beginn dieses Jahrhunderts mit der Optimierung von Selektionsverfahren, und zwar sowohl im Bildungs- wie im Beschäftigungssystem. Dabei wurden insbesondere auch die in praktischer Hinsicht zentralen Kriterien des Nutzens (oder der Nützlichkeit) und der Fairness reflektiert. Es zeigte sich dabei, dass beides zentral von der prognostischen Validität des Verfahrens abhängt. Aus dieser Sicht müsste man somit in erster Linie diese prognostische Validität optimieren. Empirische Untersuchungen zur Akzeptanz von Selektionsverfahren zeigen demgegenüber eine Präferenz für Methoden, deren Validität – und damit auch deren Nutzen und Fairness – problematisch ist. Dieser Widerspruch scheint auch in der aktuellen Diskussion um die Selektion für das Medizinstudium von Bedeutung zu sein. Im folgenden soll daher die damit angesprochene Problematik und ihr Hintergrund etwas illustriert werden. Dabei sind auch einige begriffliche Erörterungen notwendig.

Das im gegebenen Zusammenhang relevante Selektionsproblem entsteht dadurch, dass die Grundmenge der Bewerberinnen und Bewerber die Zahl der verfügbaren Ausbildungsplätze übersteigt und andere Massnahmen zur Behebung dieses Missverhältnisses durch einen politischen Entscheid ausgeschlossen werden. Aus der Grundmenge der Studieninteressenten ist daher eine Teilmenge Zugelassener auszuwählen. Sieht man – in einem ersten Verständnis – die Aufgabe eines Selektionsverfahrens nur in der Lösung genau dieses Auswahlproblems, ist jedes Verfahren nützlich, durch das die gewünschte numerische Reduktion bewerkstelligt werden kann. Und fair wäre ein solches Verfahren, wenn jeder Bewerber – unabhängig von Geschlecht, sozialer und regionaler Herkunft usw. – dieselbe Chance hat, zugelassen (oder abgelehnt) zu werden.[1] Denkt man diese Sichtweise konsequent zu Ende, ergibt sich, dass ein Losverfahren sowohl hinsichtlich Kosten-Nutzen-Verhältnis wie auch Fairness allen denkbaren Alternativen deutlich überlegen wäre.

Bereits aus Alltagssicht dürfte ein solches Zufallsverfahren aber wenig befriedigen. Sinnvoller scheint ein Selektionsverfahren zu sein, das weiterge-

---

[1] Diese Vorstellung von Fairness wird oft als "Modell der proportionalen Repräsentation" bezeichnet. Weitergehende Ausführungen zu den Konzepten Fairness und Nutzen finden sich z.B. im Buch von Amelang & Zielinski (1994, S. 130ff und S. 277f.), das auch andere testpsychologische Grundkonzepte erläutert.

henden Ansprüchen genügt: Es sollte jene Bewerber systematisch bevorzugen, die eine höhere Erfolgschance haben, und damit die Erfolgsquote in der Gruppe der Zugelassenen gegenüber derjenigen in einer Zufallsauswahl erhöhen. Oder anders formuliert: Die Anzahl der mit jedem Selektionsverfahren verbundenen Fehlentscheide, konkret: die "falsch-positiven" Entscheide (= Zulassung später Nicht-Erfolgreicher) und die "falsch-negativen" Entscheide (= Ablehung potentiell Erfolgreicher), sollte möglichst tief gehalten werden. Die Anzahl solcher Fehlentscheide ist aber direkt von der prognostischen Validität des Verfahrens abhängig, das heisst von der (nur) auf statistischem Wege feststellbaren Güte oder Treffsicherheit, mit der ein Verfahren die Erfolgswahrscheinlichkeit eines Individuums abzuschätzen erlaubt. Der Nutzen eines in diesem Sinne möglichst validen Verfahrens für die Bildungsinstitution liegt auf der Hand: Es können nicht-zielführende Ausbildungskosten eingespart werden. Aus der Sicht der betroffenen Individuen ist die Nutzenfrage zwar komplexer, weil Fallunterscheidungen notwendig sind. Eigentlich problematisch ist sie aber "nur" im Fall eines falsch-negativen Entscheids (wobei daran zu erinnern ist, dass die Anzahl solcher Fehlentscheide mit zunehmender Validität des Verfahrens abnehmen). Auch das Problem der Fairness stellt sich differenzierter dar als oben: Fair wäre ein Verfahren, das sich ausschliesslich an der Erfolgschance eines Individuums orientiert und zwar für alle Individuen – unabhängig von ihrer Gruppenzugehörigkeit – gleichermassen. Oder anders gesagt: Keine soziale Gruppe sollte durch Fehlentscheide besonders betroffen werden, womit wieder der Problemkreis der prognostischen Validität angesprochen ist.[2]

Nützliche und faire Selektion in diesem zweiten (und substantielleren Sinne) setzt somit ein Verfahren voraus, das möglichst valide ist. Die Ergebnisse jahrzehntelanger Forschung zur Konstruktion und Bewährung von Selektionsverfahren in den verschiedenartigsten Anwendungsbereichen lehrt, dass psychometrisch konstruierte und explizit an einem Validitätskriterium optimierte Tests in dieser Beziehung allen anderen Verfahren deutlich überlegen sind.[3] Im deutschsprachigem Raum erfüllt im Zusammenhang mit dem hier relevanten Selektionsproblem einzig der in Deutschland entwickelte "Test für medizinische Studiengänge" (TMS) diese Bedingung. Auf dem bisher geschilderten Hintergrund handelt es sich dabei also um dasjenige Verfahren, das am ehesten

---

[2] Dieses Verständnis von Fairness wird "Modell einer fairen Vorhersage" genannt. Es kann durchaus im Widerspruch zum weiter oben skizzierten "Modell einer proportionalen Repräsentation" stehen. Wie sich im übrigen bei einer genaueren Analyse herausstellt, ist es – unabhängig vom konkreten Selektionsverfahren – nicht immer realisierbar (siehe z.B. die Ueberlegungen in Hartigan & Wigdor, 1989).

[3] Eine entsprechende Uebersicht findet sich z.B. in Schuler & Funke (1989) oder Smith & George (1992).

eine nützliche und faire Selektion für das Medizinstudium auch in der Schweiz erlaubt, wobei natürlich weitere Kontrolluntersuchungen notwendig sind.[4]

Die aktuelle Diskussion in der Schweiz zeigt jedoch, dass die eben formulierte Auffassung in weiten Kreisen nicht geteilt wird. In einem Kanton wurde sogar per Volksabstimmung ein Spitalpraktikum als Alternative in Aussicht genommen, ein Selektionsverfahren, das mit Sicherheit einer expliziten, empirisch gestützten Nutzen- und Fairnessanalyse im obigen Sinne nicht standhalten würde. Die Akzeptanz von Selektionsverfahren scheint somit von anderen Kriterien abzuhängen, als sie bisher zugrundegelegt wurden. Empirische Untersuchungen zu dieser Akzeptanz sind zwar noch selten. Sie lassen aber doch einige Schlüsse zu.

Fruhner, Schuler, Funke & Moser (1991) haben ca. 1000 Studierende befragt, um deren Präferenzen für acht verschiedene (Personal-) Selektionsverfahren festzustellen. Als klarer Spitzenreiter erwies sich dabei das Interview (Vorstellungsgespräch). Psychologische Tests kamen deutlich schlechter weg, immerhin aber noch besser als das Losverfahren, das (zusammen mit graphologischen Gutachten) am Ende der Rangliste stand. Interessant ist, was hinter diesen Präferenzen steht, wobei bei dieser Fragestellung nur Gespräch und

---

[4] Hier ist vielleicht noch eine Präziserung am Platz: Am genannten Test wird oft kritisiert, dass er nicht am Berufserfolg sondern am Ausbildungserfolg orientiert ist. Aus der oben skizzierten Perspektive ist aber diese Konzentration auf den Ausbildungserfolg der einzige praktikable und psychologisch verantwortbare Weg: Jedes Selektionsverfahren, das sich an Kriterien des Verhaltens im Beruf zu orientieren versuchte, hätte mit Sicherheit eine deutlich geringere Validität. Gründe dafür sind unter anderem: 1.) Bei einem Studium, das auf sehr verschiedenartige Berufstätigkeiten vorbereitet, ist es äusserst schwierig, für alle künftigen Tätigkeiten gleichermassen relevante Kriterien auszumachen. 2.) Diese Kriterien wären zudem im *ärztlichen Verhalten* zu suchen, also nicht im rein kognitiven Bereich, sondern im sogenannten Persönlichkeitsbereich, dessen valide Erfassung in einer Selektionssituation weniger gut möglich ist. 3.) Generell gilt zudem, dass die Validität mit zunehmendem Zeitraum zwischen Selektion und zu prognostizierendem Sachverhalt sinkt, da die menschliche Entwicklung über längere Zeiträume sehr schwer vorauszusagen ist. Diese Gründe sprechen dafür, dass bei einem Zulassungsverfahren, das sich an Kriterien des persönlichen Verhaltens im Rahmen der weit in der Zukunft liegenden Berufstätigkeit orientieren würde, eine grosse Zahl von Fehlentscheidungen in Kauf zu nehmen wäre, mit den entsprechenden negativen Konsequenzen für Nutzen und Fairness des Verfahrens. Das an sich berechtigte Anliegen des Einbezugs von Kriterien der Berufstätigkeit sollte daher im gegebenen Fall nicht mit dem Problem der Zulassung zur Ausbildung verknüpft werden, sondern mit der Ausbildung selber und der darin ja weiterhin betriebenen Selektion.

Tests einander gegenübergestellt wurden. Generell wurde die Gesprächssituation als deutlich positiver und transparenter eingestuft als die Testsituation und als etwas weniger belastend. Fragen bei Personen, die schon Erfahrungen mit der entsprechenden Methode hatten, zeigten zudem weitere Unterschiede zugunsten des Gesprächs: Man glaubt, dass es besser geeignet ist, die relevanten Fähigkeiten zu erfassen, und bessere Möglichkeiten bietet, das eigene Ergebnis im Verfahren aktiv selber zu beeinflussen. Entsprechend wird auch das eigene Abschneiden im Gespräch durchschnittlich besser eingeschätzt als bei Tests.

Die Akzeptanz eines Selektionsverfahrens scheint somit (unter anderem) vor allem von zwei Kriterien abzuhängen. Das eine betrifft die Transparenz bzw. den subjektiven Eindruck von der Validität des Verfahrens. In der Psychologie hat sich dafür der Begriff "Augenscheinvalidität" eingebürgert, und zwar deswegen, weil dieser Eindruck einer strengen empirischen Ueberprüfung meist nicht standhält. So ist es auch beim frei und offen geführten Interview (und davon gingen die Befragten offenbar aus), das aufgrund der Befundlage zu den prognostisch wenig validen, d.h. relativ viele Fehlurteile erzeugenden Verfahren gerechnet werden muss. Das andere Kriterium hat mit der Wahrnehmung der Möglichkeit zu tun, selber aktiv und gezielt auf das Ergebnis des Selektionsverfahrens Einfluss zu nehmen. Es geht hier offenbar um die (echte oder vermeintliche) individuelle Kontrollierbarkeit des Ausgangs des Verfahrens. Die Relevanz dieses Kriteriums zeigt sich auch in einer Studie von Latham & Finnegan (1993), die sich auf die unterschiedliche Akzeptanz (dort "Praktikabilität" genannt) von unstrukturiertem und strukturiertem Interview bezieht. Als strukturiert bezeichnet man Intervews, die – im Unterschied zu unstrukturierten Interviews – weitgehend standardisiert sind und einem vorgegebenen Fragenkatalog folgen. Dies führt zu einer klar höheren prognostischen Validität, aber auch zu einer grösseren Verwandtschaft mit einer Testsituation. Auch in dieser Untersuchung zogen die Befragten das offener gestaltete Verfahren vor, z.B. mit dem Argument, sie könnten dabei ihre Motiviertheit besser zum Ausdruck bringen (a.a.O., S. 52). Dass diese (angenommene) bessere individuelle Kontrollierbarkeit des Interviewergebnisses auch Probleme hinsichtlich Validität und Fairness mit sich bringt, ist den Befragten durchaus bewusst: Auf die Frage, bei welcher Interviewform sie sich bei einem Rekurs gegen den Entscheid der Selektionsverantwortlichen mehr Erfolg versprechen, wurde wesentlich häufiger das unstrukturierte Interview genannt.

Diese Befunde sprechen dafür, dass mit Selektionsverfahren offenbar implizit ein klarer Interessenkonflikt verbunden ist: Aus "selektionstechnologischer" Sicht ist – auch im Interesse der Betroffenen – ein Verfahren vorzuziehen, das nachgewiesenermassen prognostisch valide ist, zu möglichst wenig Fehlentscheiden führt und in objektivierbarem Sinne als nützlich und fair bezeichnet werden kann. Auf der andern Seite steht das Bedürfnis nach einem transparenten, als valide erlebbaren und möglichst im eigenen Interesse beeinflussbaren

Verfahren. Beide Sichtweisen haben eine Berechtigung. Der Konflikt scheint aber nicht leicht lösbar zu sein.

Zitierte Literatur

Amelang, M. & Zielinski, W. (1994). Psychologische Diagnostik und Intervention. Berlin: Springer.

Fruhner, R., Schuler, H., Funke, U. & Moser, K. (1991). Einige Determinanten der Bewertung von Personalauswahlverfahren. Zeitschrift für Arbeits- und Organisationspsychologie, 35, S. 170 - 178.

Hartigan, J. A. & Wigdor, A. K. (1989). Fairness in employment testing. Washington D.C.: National Academy.

Latham, G. P. & Finnegan, B. J. (1993). Perceived practicality of unstructured, patterned, and situational interviews. In H. Schuler, J. L. Farr & M. Smith (eds.), Personnel selection and assessment (pp. 41 - 55). Hillsdale NJ: Erlbaum.

Schuler, H. & Funke, U. (1989). Berufseignungsdiagnostik. In E. Roth (Hg.), Organisationspsychologie (S. 281 - 320). Enzyklopädie der Psychologie. Göttingen: Hogrefe.

Smith, M. & George, D. (1992). Selection methods. In C. L. Cooper & T. Robertson (eds.), International review of industrial and organizational psychology 1992 (pp. 55 - 97). New York: Wiley.

Der "Test des Tests" - Ergebnisse eines Probelaufs des Eignungstests in der Schweiz in deutscher und französischer Sprache

Rainer Hofer, Daniel Ruefli & Klaus-D. Hänsgen

Zentrum für Testentwicklung und Diagnostik (ZTD) Universität Fribourg

Einleitung

Der aus Deutschland adaptierte Eignungstest als Kriterium für die Zulassung zum Medizinstudium in der Schweiz ist insgesamt gesehen eine gerechte wie machbare Lösung (vgl. die Beiträge von Klieme und Trost). Ein psychologischer Test, und dazu gehört dieser Eignungstest, untersteht einer genau definierten wissenschaftlichen Überprüfung. Benachteiligungen irgendwelcher Testpersonen, die durch den Test verursacht würden, dürfen demzufolge weder erwartet noch toleriert werden. Eine Überprüfung des Tests in der Schweiz war angebracht, weil in den politischen Diskussionen um den Numerus Clausus nicht der Numerus Clausus an und für sich sondern der Test als Kriterium als unfair angeprangert wurde. Test-Unfairness müsste entweder Änderung oder sogar Ablehnung des Tests als Kriterium bedeuten.

Der Probelauf in der Schweiz

Ein Probelauf des Tests in der Schweiz (Hofer et al. 1995) 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?

Die Untersuchung fand im Collège Sainte-Croix in Fribourg statt. Es wurden deutsch- und französischsprachige Gymnasiastinnen und Gymnasiasten gewonnen, die ein Jahr vor dem Maturitätsabschluss standen. Die Schülerinnen und Schüler wurden eine Woche vor dem Testtermin über den Testablauf informiert. Die verschiedenen Aufgabentypen wurden dabei besprochen. Es wurde auch ein Test-Info (Zentrum für Testentwicklung und Diagnostik 1995)

ausgeteilt, um die Möglichkeit zu bieten, sich zu Hause auf den Test vorbereiten zu 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.

Im Vergleich zu Deutschland stand für diesen Probelauf nur ein halber Tag zur Verfügung, wehalb nur sieben der neun Untertests durchgeführt wurden. Bei der deutschsprachigen Stichprobe wurden Aufgaben eines veröffentlichen Originaltests des Tests für medizinische Studiengänge verwendet (Institut für Test- und Begabungsforschung 1990). Die Aufgaben der übersetzten Verion ins Französische (Centre pour le développment de tests et le diagnostic 1996) wurde der französischsprachigen Stichprobe vorgegeben.

Insgesamt haben 42 deutsch- und 125 französischsprachige Gymnasiastinnen und Gymnasiasten den Test mit all seinen 6 Untertests (der Konzentrationstest konnte nicht ausgewertet werden) bearbeitet. Der Anteil der Frauen betrug 45 respektive 56 Prozent. Die Testpersonen erhielten für jede korrekt bearbeitete, gewertete Aufgabe einen Punkt. Es war darum möglich, im Maximum 118 Punkte zu erzielen. Im Mittel erzielte die deutschsprachige Gruppe 55,4 bei einer Standardabweichung von 11,5 Punkten. Die entsprechenden Daten für die französischsprachige Gruppe betragen 57,3 für den Mittelwert und 12,4 Punkte für die Standardabweichung.

Der Test erfüllt die Gütekriterien

Die Messgenauigkeit des Tests und seiner Untertests wurde mittels des "CRONBACH-α-Koeffizienten" bestimmt. Es zeigte sich, dass bei der deutschsprachigen Gruppe die α-Werte um den Median 0,71 liegen. Die französischsprachige Gruppe weist für den Gesamttest einen Median von $\alpha = 0{,}65$ auf. Zu dessen tieferen Wert hat der Untertest "Fakten lernen" mit dem α-Wert von 0,46 beigetragen. Es besteht hier die Annahme, dass in diesem Untertest die Eselsbrücken, die in der deutschsprachigen Fassung für das Einprägen der Fakten vorhanden sind (Hänsgen et al. 1995), in der Übersetzung verloren gegangen sind.

Der Vergleich mit den deutschen Resultaten des Testjahrganges 1992 als Kontrollgruppe (Trost et al. 1990) mit dem α-Median für den Gesamttest von 0,72 weist darauf hin, dass in diesem Probelauf eine annähernd gleich hohe Messgenauigkeit erzielt wurde. In zwei der sechs Untertests schnitten die Schweizer Stichproben im Mittel sogar 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 entspricht 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 zu der Kontrollgruppe BRD (55%) im Mittel ein Schwierigkeitsgrad von 48% erreicht. Entsprechend der Situation 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 Maturi-

tätstyp 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, 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. Die Itemanalyse nach Mantel & Haenszel (1959) wies nur in 4 von 136 Items in der deutschsprachigen und in 3 von 136 Items in der französischsprachigen Gruppe einen Bias bezüglich des Geschlechts auf. Für beide Geschlechter wurden der Inhalt, die Diskriminationsfähigkeit, das Antwortmuster und der Schwierigkeitsgrad der einzelnen Items überprüft. Es konnten jedoch keine schlüssigen Interpretationen für diese Item-Bias gefunden werden. Deshalb muss es sich um einen statistischen Zufall handeln.

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, $p < 0{,}05$, Tabelle 1). 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 Kontrollgruppe BRD (Trost et al. 1994; Hofer et al. 1995) kann festgestellt werden, dass sich die deutschsprachige Gruppe geradezu umgekehrt verhält.

Tabelle 1: Beziehung zwischen Geschlecht bzw. Sprache einerseits und Testergebnis auf der Test- und Untertestebene andererseits

	deutschsprachige Gruppe					**französischsprachige Gruppe**				
Untertest / Test	Männer(n=23)		Frauen(n=19)			Männer(n=48)		Frauen(n=60)		
	m	**s**	**m**	**s**	**p**	**m**	**s**	**m**	**s**	**p**
Textverständnis	9,04	3,91	10,37	3,15		9,60	3,04	9,80	2,83	
Figuren lernen	6,87	2,67	8,84	2,39	< 0,05	8,04	3,61	8,90	3,24	
Fakten lernen	7,26	2,86	10,26	3,83	< 0,05	10,52	2,79	11,07	2,99	
Muster zuordnen	11,35	3,52	13,16	3,24	< 0,05	12,17	3,71	11,73	2,83	
Schlauchfiguren	9,09	3,58	9,11	3,57		9,92	3,94	7,83	3,43	< 0,05
Quant. u. form. Probl.	8,26	3,98	7,84	4,36		9,29	4,26	6,87	3,00	< 0,05
Gesamt-Test	51,87	11,52	59,58	12,01	< 0,05	59,54	14,55	56,20	10,84	

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 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 gibt nur geringe sprachspezifische Unterschiede (Tabelle 1) zwischen den beiden Schweizergruppen. In den Untertests “Figuren resp. Fakten lernen” erzielten die französischsprachige Gruppe ein besseres Ergebnis als ihre deutschsprachigen Kolleginnen und Kollegen vom Collège Sainte-Croix, wobei der Unterschied im Untertest “Fakten lernen” signifikant ist. In den übrigen Untertests war kein Unterschied in den Ergebnissen beider Geschlechter vorhanden.

Bei der Itemanalyse nach Mantel & Haenszel (1959) wiesen 6 von 136 Items einen Bias bezüglich der Sprache auf, wovon 3 aus eher sprachunabhängigen Untertests stammen. Die Überprüfung dieser Items zeigt, dass die Unterschiede allerdings eher unsystematisch sind. In den Untertests “Textverständnis”, “Muster zuordnen”, “Schlauchfiguren” und “Quantitative und formale Probleme” hat die deutschsprachige Gruppe ein leicht besseres Durchschnitts-

ergebnis 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

Der Probelauf hat bestätigt, dass einerseits die organisatorisch-technischen Voraussetzungen für die Anwendung des Tests in der Schweiz erfüllt werden können, und dass andererseits von einem fairen Verfahren ausgegangen werden kann.

1. 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 Kontrollgruppe BRD.
2. Die schweizerdeutsche und die französischsprachige Adaptation des Tests erreichen bezüglich Zuverlässigkeit und Testfairness eine der deutschen Originalfassung gut vergleichbare Testgüte. Es konnte damit auch unter Beweis gestellt werden, dass das gewählte Adaptationsverfahren angemessen ist.
3. 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.

Literatur

Centre pour le développement de tests et le diagnostic, Université de Fribourg (Suisse) en collaboration avec l’institut für Test- und Begabungsforschung, Bonn, Allemagne (ed.) (1996). Le test d’aptitudes pour les études de médecin. Adaptation française de la version originale dans son intégralité. Göttingen: Hogrefe.

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.

Hofer, Rainer; Ruefli, Daniel & Hänsgen, Klaus-Dieter (1995). Der Eignungstest für das Medizinstudium in der Schweiz - Ein Probelauf. Universität Fribourg, Zentrum für Testentwicklung und Diagnostik: Bericht 1.

Institut für Test- und Begabungsforschung (Hrsg.) (1990). Test für medizinische Studiengänge. Aktualisierte Originalversion 2. Göttingen: Hogrefe, 3. Auflage.

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.

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.