Centre pour le développement de tests et le diagnostic
à l'Institut de psychologie de l'Université de Fribourg

Diagnostic d'aptitudes et études de médecine

Rapport d'un symposium, édité par
K.-D. Hänsgen, R. Hofer et D. Ruefli

Rapport 2 (1996)

Table des matières

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

Tests d'aptitudes et études de médecine

Les 27 et 28 octobre 1995 a eu lieu à Fribourg une rencontre de psychologues, de pédagogues, de médecins et de représentants d'autres disciplines pour discuter en profondeur toutes les questions scientifiques liées à un éventuel **test d'aptitudes pour les études de médecine en Suisse**.

La manifestation a été menée en commun avec l'Institut de psychologie de l'Université de Fribourg sous l'égide de la Conférence universitaire suisse. Nous remercions le Professeur Perrez, Directeur de l'Institut de psychologie, pour son aide à réaliser l'idée à la base de la manifestation, à savoir dresser un bilan concernant le **fondement scientifique** de tests d'aptitudes et définir les **exigences** requises pour une éventuelle application du test. A la séance plénière du vendredi, suivie d'une discussion animée, a succédé le samedi un workshop (atelier-conférence) avec les hôtes internationaux.

La rencontre a été consacrée pour l'essentiel à la discussion d'expériences internationales recueillies avec des tests d'admission aux études, utilisés dans la majorité des pays industrialisés et qui y sont reconnus, pour autant qu'un numerus clausus ait été jugé nécessaire, comme méthode équitable et faisable de régulation de l'accès aux études de médecine.

R. Hambleton (University of Massachusetts, Etats-Unis) a présenté le Medical College Admission Test (MCAT), appliqué par l'Association of American Medical Colleges comme test d'entrée. Il s'agit d'un test de l'aptitude à faire des études et qui examine en particulier l'aptitude à résoudre des problèmes. Il est adapté et révisé à intervalles réguliers de manière à correspondre aux exigences requises pour réussir des études de médecine. Le deuxième grand thème d'études de la rencontre a été la question des critères sur lesquels se fonder pour la traduction d'items de test et d'examen dans d'autres langues. Il existe à ce sujet des règles, reconnues à l'échelle internationale, émanant de l'International Test Commission et qui fournissent de précieuses indications, valables aussi pour les traductions à faire en Suisse.

M. Beller (University Tel Aviv, Israèl) a rapporté sur le test d'aptitudes utilisé en Israèl pour l'admission aux études de médecine. Fait remarquable, ce test peut être passé en six langues. L'équivalence des traductions dans cinq langues à partir de l'hébreu est vérifiée continuellement. Les orateurs ont présenté des possibilités pratiques de contrôle et de correction, qui garantissent l'égalité des chances. ***Ch. Stage*** (Universität Umeå, Suède) a démontré les principes du Swedish Scholastic Aptitude Test (SweSAT), utilisé depuis 1977 et adapté lui aussi en 1991. En six sous-tests sont examinées des prestations importantes pour les études. Ce test a lieu deux fois par an et donne des indications

également sur l'aptitude à faire des études. Il est veillé en particulier, lors de l'élaboration du test, à le faire correspondre autant que possible aux exigences requises pour la réussite des études. ***G. Trost et E. Klieme*** (Institut für Test- und Begabungsforschung Bonn, Allemagne) ont pu présenter de récents résultats obtenus avec le test en usage en Allemange pour les études de médecine (Test für Medizinische Studiengänge, TMS) qui fournissent la preuve de sa très grande valeur de prévision du succès des études, valeur supérieure à celle du recours aux simples notes de maturité. Concernant le problème de l'équité, l'orateur a démontré que l'application du test fournit en soi la garantie du respect de ses conditions.

Les tests présentés ont tous été conçus de manière à ce qu'ils puissent être passés avec succès sans connaissances médicales préalables, de manière à éviter dans toute la mesure du possible que les candidats puissent s'y préparer par un entraînement. C'est une propriété importante que doit présenter le test pour être équitable du point de vue social, c'est-à-dire pour assurer l'inefficacité et, partant, l'inutilité de cours onéreux de préparation.

R. Bloch (Université de Berne) a rapporté sur les expériences faites à la suite de l'application d'une procédure de sélection complexe à l'Université de McMaster au Canada, où l'on se fonde avant tout sur des entretiens d'aptitudes. ***F. Baumann*** (Université de Genève) a parlé de la place à accorder au savoir et aux aptitudes qu'un bon médecin doit avoir à l'heure actuelle. ***U. Schallberger*** (Université de Zurich) a fait voir que l'utilité et l'équité sont les deux notions d'évaluation les plus importantes pour les procédures de sélection. A partir de réflexions théoriques, il a montré les problèmes que pose l'obligation d'accorder un traitement égal à des groupes divers de candidats et esquissé des solutions pour y parvenir.

Les représentants du Centre pour le développement de tests et le diagnostic ***(R. Hofer, D. Ruefli et K.-D. Hänsgen)*** ont présenté de premiers résultats d'un **essai pilote de passation du test d'aptitudes** en allemand et en français avec des items rédigés dans chacune de ces deux langues (l'essai a été effectué au Collège Sainte-Croix à Fribourg). Bien que les candidats fortuits ne fussent pas véritablement en situation de vouloir faire des études de médecine, le test a atteint des critères de qualité presque égaux à ceux obtenus en Allemagne. Le test différencie suffisamment, pour ce qui est de la prestation et, partant, de l'aptitude à faire des études, pour justifier le recours à cette procédure pour l'admission aux études de médecine. Comme en Allemagne, environ 50%, en moyenne, des problèmes à résoudre le sont correctement.

Les conclusions à tirer du symposium concernant l'application d'un test d'aptitudes en Suisse ont été résumées par ***N. Ischi*** (Secrétaire général de la Conférence universitaire suisse).

Dans la présente brochure ont été reproduits quelques-uns des exposés de la rencontre comme incitation à poursuivre la discussion sur l'utilité et les problèmes possibles de l'application d'un test d'aptitudes en Suisse. La publication a également pour but d'informer dans l'espoir d'accroître l'objectivité des discussions et d'amener à voir le test sans partis pris. Nous croyons qu'il n'y a pas, en définitive, de méthode de sélection meilleure qu'un test quand il devient nécessaire de décréter le numerus clausus. Une sélection fondée sur des entretiens est une solution beaucoup trop coûteuse si elle doit être pratiquée avec tous les candidates et candidats. On pourrait tout au plus concevoir des entretiens comme possibilité de nuancer les prestations fournies par les personnes situées en zone limite d'admission, en lieu et place d'une liste d'attente. Ce n'est pas en faisant effectuer un stage aux candidats que l'on parviendra à départager ceux qui se prêtent à des études et ceux qui ne s'y prêtent pas, abstraction faite de la nécessité de disposer des places de stage en nombre suffisant. On sait par expérience qu'un très petit nombre de jeunes gens désireux d'étudier la médecine renoncent à se porter candidat à la suite des impressions que leur a laissées un stage hospitalier. Les notes de maturité ne sont pas comparables, en Suisse. La valeur prédictive du test pour le succès des études est attestée scientifiquement. Un nombre considérable de pays ont fait des expériences positives avec l'utilisation de cet instrument de sélection. Il n'est pas rédhibitoire et est équitable tant du point de vue des groupes linguistiques que des sexes.

Klaus-D. Hänsgen

Directeur du CTD

Littérature:

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 Swe-SAT 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 verbal 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 pro-

cedure 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

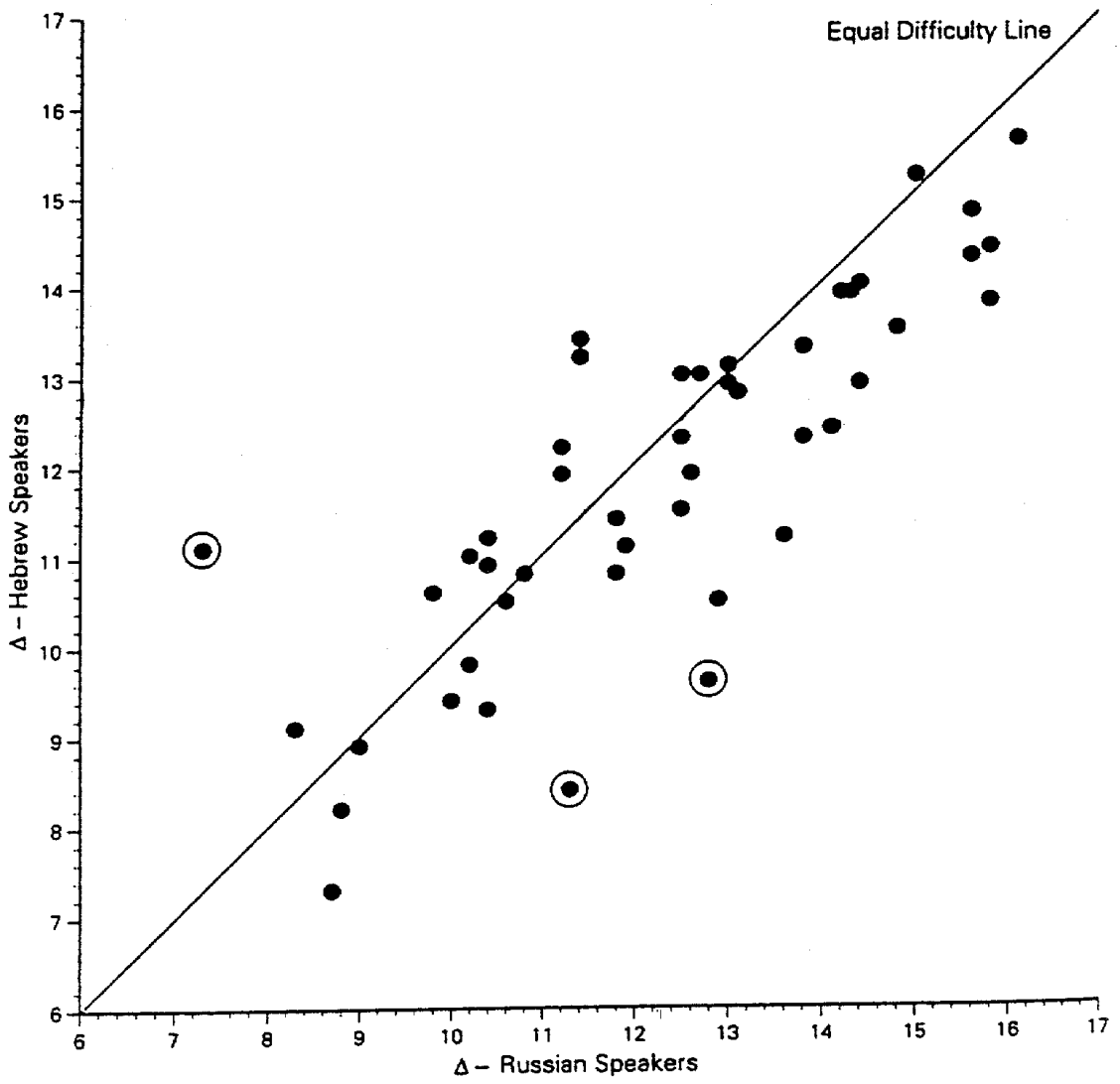
Q	44	11.16	11.15	2.21	1.91	.57	.67	.10	.13	.92
2										
V	47	12.37	11.71	1.98	1.91	.41	.42	.12	.09	.70
Q	44	11.74	11.25	2.09	2.15	.54	.55	.10	.09	.96
3										
V	50	11.67	10.87	2.16	2.04	.46	.44	.10	.10	.81
Q	44	10.52	11.87	2.10	1.66	.55	.53	.15	.12	.92

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 () 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.

Prévisions de succès dans les études de médecine - de la validité du test pour les filières médicales et d'autres instruments de sélection

Eckhard Klieme

Institut de recherche en matière de tests et d'aptitudes, Bonn (Allemagne)

1. Le test pour les études de médecine (TMS) et son utilisation pour l'admission aux études de médecine en Allemagne

Le test pour les études de médecine (TMS) a été développé à la fin des années septante à l'Institut de recherche en matière de tests et d'aptitudes et mis à l'épreuve sur une large échelle durant la période allant de 1980 à 1985. Depuis 1986, tous les candidats aux études de médecine humaine, médecine dentaire et médecine vétérinaire (à quelques rares exceptions près) sont tenus, en République fédérale d'Allemagne, de participer au test.

Pour chaque session de test est développée une nouvelle version du TMS. La composition du test reste cependant la même: divisé en plusieurs sous-tests, il comprend au total 184 items, que les candidats doivent exécuter selon le principe des questions à réponses multiples; à cela s'ajoute une procédure spéciale visant à mesurer la capacité à travailler avec soin et concentration. Les quatre groupes principaux d'items (ils prennent à eux seuls quatre des cinq heures de temps accordées pour faire le test) visent à discerner la capacité à penser logiquement dans des contextes médicaux et scientifiques, par exemple à interpréter des graphiques et des tableaux, à comprendre des passages de textes d'une certaine longueur et à résoudre des problèmes quantitatifs et formels. Le test comprend un deuxième grand groupe d'items qui vérifient différents aspects du traitement d'informations visuelles, par exemple la capacité à se représenter les choses dans l'espace, la précision de l'appréhension ainsi que la concentration et le soin avec lesquels sont reconnus des signes visuels. Une troisième partie du test est destinée à évaluer la capacité de mémorisation (tant d'informations linguistiques que d'informations visuelles).

Le test a pour but de déterminer l'aptitude à faire des *études*; ses items visent à reproduire les exigences fondamentales des études. Il serait intéressant de savoir dans quelle mesure le succès au test et durant les études correspond au succès que le médecin aura plus tard dans sa profession. Etant donné le nombre élevé de champs professionnels médicaux et faute de critères sûrs, il n'est cependant guère possible d'estimer "le succès professionnel" de manière empirique.

Lors de l'admission aux études de médecine, il est procédé de la manière suivante depuis le semestre d'hiver 1986/87: 45 pour cent des places d'études disponibles sont attribués selon la combinaison "note moyenne de maturité et prestation au test", c'est-à-dire que dans le pourcentage "maturité/test" entrent

les candidats qui présentent les meilleures "valeurs" (calculées selon la formule "rapport 55:45 somme pondérée du résultat de la maturité et résultat du test"). 10 autres pour cent des places vont aux candidats qui (après calculation du quota abitur/test) ont obtenu les meilleurs résultats de TMS. 20 pour cent des places d'études sont accordés après un temps d'attente, 15 pour cent selon le résultat d'entretiens de sélection dans les universités. Les places d'études restantes (environ 10 pour cent) demeurent réservées pour des groupes déterminés de candidats (étrangers, "cas de rigueur", etc.). Le ou la candidat(e) qui n'a été pris en considération dans aucun de ces groupes peut présenter à nouveau sa candidature aussi souvent qu'il le veut lors de semestres ulté-rieurs.

2. Questions et base des données de l'étude de validité

L'Institut de recherche en matière de tests et d'aptitudes n'a pas seulement pour tâche de développer le test; il doit également l'évaluer. L'une des questions les plus importantes de l'évaluation est de savoir quelle sera la qualité des prestations fournies durant les études par les étudiants admis selon les différents critères décrits ci-avant dans les groupes (“quotas”). Par exemple, les étudiants admis sur la base de bonnes prestations de maturité et de test réussissent-ils mieux leurs examens que ceux qui ont reçu leur place d'études après un long temps d'attente? L'étude qui sera présentée ici devrait en outre indiquer dans quelle mesure les prestations fournies aux examens oraux et aux examens écrits peuvent de manière générale être prédites à partir de prestations scolaires et de tests.

Une enquête menée sur une vaste échelle a réuni des données sur le déroulement des études de 28 000 personnes. Ces personnes étaient celles qui ont passé le test en 1986 ou en 1987 (c'est-à-dire après que le TMS fut devenu obligatoire), qui ont été admises l'une des années suivantes à commencer des études de médecine (humaine) et qui se sont présentées jusqu'au 31.12.1992 (jour repère) à l'examen propédeutique de médecine, lequel peut être passé au plus tôt après quatre semestres d'études.

Pour chacune de ces personnes ont été relevées les données suivantes: la note moyenne obtenue à la maturité, les prestations fournies au test (test entier et sous-tests), le critère de l'admission, les prestations fournies à l'examen propédeutique (examen oral et branches de l'examen écrit, consistant en des questions à réponses multiples), le cas échéant, aussi les prestations fournies lors d'examens de répétition et la durée des études jusqu'à l'examen.

3. Dépendance entre succès aux études et critères de sélection

Le terme de réussite s'applique ici aux personnes qui ont réussi du premier coup l'examen propédeutique après une durée minimale d'études, c'est-à-dire après quatre semestres d'études. Des 28 000 étudiantes et étudiants examinés, 66 pour cent ont réussi de cette manière l'examen propédeutique, considéré comme l'obstacle le plus important des études. Parmi ceux qui ont été admis dans le groupe dont la note de maturité et de test combinée était suffisante pour se présenter au test, le taux de réussite a été même de 80 pour cent; les candidats admis dans le groupe dont le résultat au test a été suffisant ont présenté un taux de réussite de 62 pour cent; dans les cas d'admission après un temps d'attente, le taux de réussite a été de 45 pour cent et parmi ceux qui doivent leur admission à un entretien de sélection, le taux de réussite a été de 49 pour cent.

Les différences constatées entre les groupes admis selon des critères différents sont très grandes. A noter cependant que ces critères ne sont pas indépendants les uns des autres. Par exemple, n'ont droit à un entretien de sélection que les candidats dont les prestations de maturité et de test ne suffisent pas pour une admission. Etant donné cette "présélection" on peut argeur en faveur de l'entretien que les taux de réussite des personnes ayant passé un tel entretien sont encore plus élevés que ceux des personnes admises après un temps d'attente.

Seuls des calculs découlant de simulations permettent de développer des mesures de comparaison "absolues". Ainsi, nous avons pu estimer qu'avec une sélection fortuite, c'est-à-dire en cas d'attribution des places d'études par tirage au sort, quelque 48 pour cent des candidats admis au sens défini ci-avant auraient réussi l'examen propédeutique. Il apparaît dès lors que la procédure appliquée, en particulier la prise en considération de prestations de tests et de notes de maturité, augmente le taux de succès de presque 20 points de pourcent. (Pour plus d'informations concernant les résultats de simulations, lire l'exposé tenu par Günter Trost lors du symposium.)

4. Lien entre les prestations de test, la note de maturité, d'une part, et les prestations fournies lors d'examens en médecine, d'autre part

Le rapport entre le résultat au TMS d'une part et la note globale de l'examen propédeutique d'autre part peut être décrit assez bien par une équation linéaire. L'étroitesse de ce rapport est traduite en un chiffre appelé coefficient de validité, qui, en tant que mesure de corrélation, peut être une valeur se situant entre 0 et 1. Dans notre étude, le coefficient a été de 0,45. En regard de

résultats internationaux, qui font tous état de coefficients de validité variant entre 0,30 et 0,60, la validité du TMS peut être qualifiée de haute si l'on considère que le succès aux études est pronostiqué ici sur un laps de deux ans au moins, ce qui est un temps assez long.

La corrélation entre la note moyenne de maturité et la note globale de l'examen propédeutique peut être chiffrée à 0,47. Dès lors ce rapport est lui aussi relativement étroit. Le fait que la combinaison du résultat du test et de la note de fin de la scolarité, ce que nous appelons la "valeur", qui se situe à 0,54, présente une validité nettement plus élevée, revêt une importance pratique très grande. Manifestement, le test et la note scolaire appréhendent des aspects différents des aptitudes aux études et leur combinaison améliore nettement la qualité des prévisions de succès.

Il est intéressant en outre de considérer séparément la partie orale et la partie écrite de l'examen propédeutique. La prestation orale de l'examen est plus difficile à pronostiquer que la partie écrite, comme le montrent les coefficients de respectivement 0,36 et 0,57 pour la corrélation avec la "valeur". Il n'est pas rare que ce phénomène soit dû à la précision moins grande avec laquelle sont mesurées les prestations des examens oraux; certains signes semblent indiquer qu'il y aurait un "effet de méthodes". La note obtenue à l'examen oral dépend un peu plus étroitement de la note de maturité (notoirement composée elle aussi de prestations d'examens oraux) que du résultat du test; pour l'examen écrit, c'est l'inverse.

Des enquêtes antérieures menées dans les années huitante par petits échantillons ont abouti à des résultats presque identiques. Nous avons pu constater que lors d'examens ultérieurs passés durant les années d'études cliniques il y avait toujours, après cinq à six années d'études, un rapport considérable entre les critères d'admission et le résultat des examens. Les coefficents de corrélation sont inférieurs d'environ 0,10, le modèle est identique.

Des enquêtes toutes récentes indiquent que le succès obtenu durant les études de médecine dentaire et les études de médecine vétérinaire peut pratiquement être tout aussi bien pronostiqué que pour les études de médecine.

5. Un modèle de structure permettant de prédire le succès qui sera obtenu dans les curricula de médecine

Pour terminer, nous procéderons, à l'intérieur du test pour les études de médecine, à une distinction entre différentes parties du test. Nous utiliserons à cet effet un modèle de mise en équation des structures, dont les bases statistiques ne peuvent pas être expliquées ici (cf. à ce sujet les rapports de travail de l'Institut de recherche en matière de tests et d'aptitudes). L'idée de base est de résumer les différents sous-tests du TMS en ce qu'il a été convenu d'appeler

des facteurs, qui représentent des dimensions de prestation quasi idéales, dénuées d'erreurs de mesure. Comme nous l'avons déjà mentionné au chapitre 1, les facteurs peuvent être classés en (1) "être capable de conduire un raisonnement et de parvenir à des conclusions" (2) "être capable de traiter l'information de manière visuelle" et (3) "être capable de mémoriser" (3). Les six branches testées dans la partie écrite de l'examen propédeutique de médecine (physique, chimie/biochimie, biologie, physiologie, anatomie et psychologie/sociologie médicales) constituent un seul facteur à l'intérieur duquel on ne peut pas distinguer de dimensions spécifiques de prestation.

Ce modèle, qui présente une bonne adaptation aux données à disposition, a débouché sur les coefficients suivants pour les corrélations avec le facteur de l'examen propédeutique en médecine: 0,62 pour le 1er facteur de TMS, 0,23 pour le deuxième et 0,32 pour le troisième facteur. La valeur prédictive du TMS repose, d'après ces résultats, de manière prépondérante sur l'appréhension des capacités à raisonner logiquement dans des contextes de médecine et de sciences naturelles. Le concept du test, qui s'appuie très fortement sur la "simulation" d'exigences complexes des études (par exemple de la compréhension de textes longs, de l'interprétation de graphiques et de tableaux, du maniement de chiffres, d'unités et de formules), se révèle, compte tenu de ces données, utilisable aux fins visées.

La valeur de résultats de test en comparaison de notes scolaires comme critères de sélection: effet d'ascenseur, effet de filtre, effets coûts/rendement et répercussion sur l'équité de l'admission

Günter Trost

Institut de recherche en matière de tests et d'aptitudes, Bonn (Allemagne)

Pour juger de l'utilité de tests de capacité aux études lors de l'admission à des études déterminées, il y a une série d'aspects à considérer. L'un des aspects les plus importants est la valeur prédictive de tels procédés de tests concernant le succès dans les études concernées; les résultats d'enquêtes empiriques concernant cette question ont été indiqués dans les exposés de Michal Beller et d'Eckhard Klieme. D'autres questions fondamentales ont pour but (a) d'établir la relation entre les prestations fournies au test et les prestations scolaires ainsi que les conséquences découlant de l'étroitesse de ce rapport quand l'un des critères de sélection est remplacé par un autre, (b) l'influence que le résultat d'un test de capacité aux études exerce sur le choix des études des participants au test, (c) le rapport coûts/rendement du développement et de l'utilisation d'un test de capacité aux études et (d) l'équité d'un procédé de sélection basé sur des prestations fournies lors d'un test comparée à l'équité par exemple d'une

admission sur la seule base de la note scolaire terminale. Le présent exposé contient des réponses aux questions telles qu'elles ont émané des enquêtes d'accompagnement du test pour l'admission aux études de médecine en République fédérale d'Allemagne.

1. Combien étroit est le rapport entre la prestation fournie au test pour l'admission aux études de médecine et la prestation scolaire et quel est son effet?

L'indice d'étroitesse du rapport entre le résultat global obtenu au test pour l'admission aux études de médecine (TMS) et la note moyenne obtenue dans le certificat de la maturité est de 0,40. (Il s'agit d'un coefficient de corrélation qui peut varier sur une échelle allant de 0,00 - pas de lien de quelque nature que ce soit - à 1,00 - lien absolu.) Un indice de cet ordre désigne un lien moyen. Cela signifie que le test permet d'évaluer surtout des capacités qui ne se reflètent pas dans l'appréciation scolaire des prestations, mais qui sont importantes pour le succès dans les études de médecine, comme l'attestent les résultats des contrôles d'efficacité (voir l'exposé d'Eckart Klieme).

Si l'on utilise une combinaison de la note moyenne de la maturité et du résultat du test comme critère d'admission aux études de médecine, comme c'est le cas en République fédérale d'Allemagne, cette combinaison étant appelée "quota principal d'admission", on constate le phénomène suivant: environ 30 pour cent des candidats admis doivent cette admission à leur bonne prestation au test; ils n'auraient pas obtenu de place d'études si le choix avait été effectué seulement sur la base de la note moyenne de la maturité (selon la pondération de la note moyenne de maturité et du résulat du test, certains candidats obtiennent une position plus élevée sur l'échelle des points, entraînant une position plus basse pour les autres, ou "effet d'ascenseur").

2. Le résultat du test influe-t-il sur la décision des personnes intéressées à faire des études de médecine, ou en d'autres termes, le test a-t-il pour incidence qu'elles se portent effectivement candidates à l'admission à des études de médecine ou non?

Toutes les personnes qui ont acquis la maturité générale ou qui sont élèves du degré 13 du gymnase ont la possibilité, en République fédérale d'Allemagne, de prendre part au TMS. La participation est gratuite. Les personnes qui se portent candidates à l'obtention d'une place d'études dans les études de médecine doivent auparavant avoir fait le test.

Une étude longitudinale s'étendant sur un laps de quatre ans a porté sur le comportement en matière de candidature de toutes les personnes qui avaient pris part au TMS à l'automne 1986. Il est apparu ce qui suit:

(a) Durant les quatre ans qui ont suivi la participation au test, 34 pour cent des personnes qui avaient effectué le test ne se sont pas (humaine, dentaire ou vétérinaire).

(b) En moyenne, la prestation de test de ceux et de celles qui ne se sont pas portés candidats à des études de médecine par la suite a été nettement inférieure à la prestation au test de ceux et de celles qui se sont portés candidats à l'obtention d'une place d'études en médecine. (En revanche, les prestations scolaires des personnes qui se sont portées candidates à l'obtention d'une place d'études en médecine par la suite se sont révélées être, en moyenne, seulement un peu meilleures que les prestations scolaires des personnes qui ne se sont pas portées candidates à l'admission aux études de médecine par la suite.)

Le test exerce entre autres choses un effet souhaité de filtre. Le résultat du test influe sur la décision des participants de se porter ou non candidats à l'admission à l'un des curricula de médecin. Cette "auto-sélection" décharge la procédure institutionnalisée de sélection.

3. Quel est le rapport coûts/rendement lors de l'utilisation d'un test de capacité aux études du genre du TMS?

Les coûts d'une procédure de test telle que le TMS sont faciles à chiffrer. En République fédérale d'Allemagne, ils se montent, convertis en francs suisses, à environ 2 millions par année. Par contre, l'utilité d'une telle procédure de test est en partie de nature immatérielle et ne peut dès lors que difficilement être quantifiée en sommes d'argent.

Néamoins, l'un des aspects d'utilité entrant en considération peut être converti - dans le contexte allemand de sélection et d'études - en économies effectives, comme le montrera l'exemple ci-après. En se fondant sur les données provenant du contrôle d'efficacité, relatives à environ 28 000 étudiants en médecine (voir l'exposé de Eckart Klieme) on parvient par des calculs-types à déterminer le pourcentage de personnes qui réussiraient du premier coup l'examen propédeutique de médecine, quel que soit le nombre de semestres accomplis, si les places d'études étaient attribuées fortuitement, c'est-à-dire sans aucune sélection systématique. Le critère de succès construit ici reçoit une définition plus large que pour le calcul-type mentionné dans l'exposé de E. KLIEME. Ce "taux de base" est de 69 pour cent. Si l'on sélectionnait non pas de cette manière mais en se fondant uniquement sur les résultats obtenus au test pour l'admission aux curricula de médecine, le taux de succès, parmi les

personnes admises, serait inférieur à 90 pour cent. La part de ceux qui étudient au moins un semestre de plus parce qu'ils doivent répéter l'examen propédeutique en médecine se réduit de 21 points de pour-cent; cela correspond, en République fédérale d'Allemagne, à 1 650 personnes par année.

Les coûts d'études complètes de médecine à payer par le contribuable sont chiffrés en République fédérale d'Allemagne à un montant qui, converti en francs suisses, est d'environ 320 000 francs (il n'est pas perçu de taxes d'études); cela correspond, par place d'études et semestre, à un montant de 27 000 francs suisses. L'utilisation des résultats de test comme critère de sélection en lieu et place d'une sélection fortuite et l'amélioration du taux de succès obtenue par ce système de sélection permettent, en République fédérale d'Allemagne, d'économiser un montant qui, converti en francs suisses, est d'un peu moins de 45 millions de francs en coûts d'études, contre 2 millions de francs suisses pour le développement et la passation du test. Si l'on n'établit que l'accroissement du taux de succès obtenu en ajoutant le résultat du test à la note moyenne de la maturité comme critère de sélection, on parvient à une amélioration de quatre points de pour-cent; cette différence correspond à une économie annuelle de coûts d'études de l'ordre de 8,5 millions de francs suisses.

4. Quelle est l'équité d'une procédure de sélection basée sur la prestation au test par rapport à une procédure de sélection fondée sur la note moyenne de la maturité?

La question de l'équité de la procédure d'admission aux études de médecine sera discutée dans l'optique de groupes de candidats, l'un étant les candidats masculins et l'autre les candidats féminins.

Pour commencer, quelques faits:

La note moyenne de la maturité des candidats féminins à des places d'études dans les études de médecine est, en moyenne, très légèrement supérieure à la note de maturité des candidats masculins (différence: de 0,12 unité de note sur l'échelle de 1,0 - meilleur résultat - à 4,3 - résultat le plus faible -; l'écart-type est pour les deux groupes de 0,6 unité de note.)

La valeur totale telle qu'elle est calculée pour le test d'admission aux études de médecine se situe en moyenne, chez les candidats masculins, un peu au-dessus de la valeur totale des candidats féminins (différence: 2,2 points-types sur une échelle allant de 70 à 130 points; la valeur moyenne de tous les participants au test est de 100; l'écart moyen est pour les deux groupes de 10 points-types).

A l'étude de ces données, on serait tenté de conclure hâtivement qu'une admission sur la seule base de la note moyenne obtenue à la maturité désavantage les hommes, alors qu'une admission sur la seule base du test désavantage les femmes. Ces conclusions reposent sur une conception de l'équité de nature suivante: “Une procédure de sélection est équitable si la part des membres de

groupes partiels définis par rapport aux personnes sélectionnées est égale à la part de ces groupes partiels par rapport au groupe total des candidats.” Cette définition omet toutefois les aptitudes effectives des membres des groupes partiels, c'est-à-dire le succès aux études auquel on peut s'attendre de leur part.

En psychologie et dans les sciences sociales on accorde de ce fait la préférence à d'autres définitions de l'équité, qui tiennent compte des aptitudes des membres des groupes partiels concernés. Une telle définition est formulée comme suit: “Une procédure de sélection est équitable par rapport à des groupes déterminés de candidats si l'on donne aux groupes partiels qui ont les mêmes perspectives de succès les mêmes chances d'admission.” A l'étude des prestations des femmes et des hommes (non sélectionnés sur la base de notes scolaires ou de résultats de test) il apparaît que lors de l'examen propédeutique de médecine, par exemple, les hommes obtiennent en moyenne une valeur-type (également sur l'échelle de 70 à 130 points; écart-type 9 points) supérieure de presque trois points à celle obtenue par les femmes.

La procédure de sélection ne répond qu'insuffisamment à ces perspectives meilleures des hommes. En effet, si l'on met en rapport la prestation fournie à l'école, ou celle fournie au test avec la prestation fournie aux études, on parvient aux constats suivants:

- Les candidats masculins sont désavantagés dans les deux cas par rapport à leurs prestations aux études, c'est-à-dire tant si l'admission se fait sur la base de la seule note de maturité que si elle se fait sur la base des seuls résultats au test.
- L'ampleur de l'inéquité est cependant nettement moindre si le résultat du test est utilisé comme critère de sélection.

Utilité, équité, validité et acceptabilité de procédures de sélection

Urs Schallberger

Université de Zurich, Institut de psychologie

La psychologie appliquée étudie depuis ses origines, c'est-à-dire depuis le début de ce siècle, les possibilités d'optimisation de procédures de sélection, et cela tant dans le système d'éducation que dans le système d'emploi. Ces études portent également sur les critères cruciaux de l'utilité et de l'équité, fondamentaux en particulier sous l'angle pratique. Il est apparu à ce sujet que les deux dépendent fortement de la valeur prédictive de la procédure. Vu ce constat, il faudrait dès lors en premier lieu optimiser la valeur prédictive. Des enquêtes empiriques concernant l'acceptabilité de procédures de sélection indiquent en revanche une préférence pour des méthodes dont la valeur - et, partant également leur utilité et équité - est problématique. Cette contradiction semble revêtir de l'importance également dans la discussion actuelle relative à la sélection pour les études de médecine. Dans les pages qui suivent nous tenterons d'illustrer quelque peu cette problématique et d'esquisser la toile de fond sur laquelle elle s'inscrit. Quelques indications notionnelles sont nécessaires à cet effet.

Le problème de sélection qui se pose dans le contexte donné provient du fait que la quantité de base des candidates et des candidats excède le nombre des places de formation à disposition et que d'autres mesures visant à supprimer cet écart par une décision politique sont exclues. C'est pourquoi il faut procéder, parmi la quantité de base des personnes intéressées à faire des études de médecine, à une sélection pour réduire la quantité de base à une quantité partielle de personnes admises. Si l'on conçoit, à première vue, que le but d'une procédure de sélection est uniquement de trouver une solution à ce seul problème de sélection, toute procédure est utile qui débouche sur la réduction numérique souhaitée. Et une telle procédure serait équitable si chaque candidat - indépendamment de son sexe et de sa provenance sociale et régionale, etc. - avait les même chances d'être admis (ou d'être écarté[1]). Ce raisonnement poussé jusqu'au bout fait apparaître qu'une procédure par tirage au sort serait nettement supérieure à toutes les alternatives concevables, et cela tant du point de vue du rapport coût/rendement que du point de vue de l'équité.

---

1 Cette conception de l'équité est souvent appelée "modèle de la représentation proportionnelle". Pour de plus amples considérations relatives aux notions d'équité et d'utilité, lire l'ouvrage d'Amelang & Zielinski (1994), p. 130 et suiv. et 277 et suiv., qui traite également d'autres notions fondamentales de psychologie de tests.

Mais une telle procédure sera sans doute jugée d'emblée peu satisfaisante. Une procédure de sélection qui satisfasse à de plus amples exigences semblera plus judicieuse: elle devrait être de nature à accorder systématiquement la préférence aux candidats qui ont une plus grande chance de succès et qui, partant, augmentent le taux de succès dans le groupe des personnes admises par rapport au taux de succès enregistré dans un groupe de personnes constitué au hasard. Ou, en d'autres termes: le nombre des fausses décisions lié à chaque procédure de sélection, c'est-à-dire les décisions "positives et pourtant fausses" (= admission de personnes qui ne réussiront pas leurs examens par la suite) et les décisions "négatives et pourtant fausses" (= refus de bons étudiants potentiels), devrait être aussi bas que possible. Or le nombre de telles décisions fausses dépend directement de la valeur prédictive de la procédure, c'est-à-dire de sa qualité ou de la sûreté de ses choix, ou encore de sa capacité à estimer les probabilités de succès des candidats aux examens, constatables (uniquement) par la voie de la statistique. L'utilité, pour une institution de formation, d'une procédure de bonne valeur au sens indiqué est évidente: elle permet d'économiser des coûts de formation ne menant pas au but. Dans l'optique des individus concernés, la question de l'utilité est évidemment plus complexe parce qu'il faut prendre des décisions au cas par cas. Mais elle n'est en fait problématique "que" dans le cas d'une décision négative et pourtant fausse (à noter toutefois que le nombre de fausses décisions de ce type se réduisent en proportion de la valeur de la procédure). Le problème de l'équité se pose lui aussi en termes plus différenciés que ci-avant: serait équitable une procédure axée seulement sur les chances de succès d'un individu et cela de manière égale pour tous les individus, indépendamment de leur appartenance à un groupe. Ou, autrement dit: aucun groupe social ne devrait être touché plus spécialement par des décisions fausses, et l'on rejoint là le problème de la valeur prédictive.[2]

Une sélection utile et équitable dans ce deuxième sens (plus substantiel) présuppose dès lors une procédure qui soit aussi valable que possible. Les résultats de recherches menées durant de longues années pour la construction et la mise à l'épreuve de procédures de sélection dans les domaines d'application les plus divers montrent que des tests construits psychométriquement et optimisés explicitement sur la base d'un critère de validité sont nettement supérieurs à toutes les autres procédures à cet égard.[3] Dans les régions germanophones seul le test développé en Allemagne pour les études de médecine, le "Test für medi-

---

2 Cette acception de l'équité est appelé "modèle de la prédiction équitable". Il peut parfaitement être en contradiction avec le "modèle de la représentation proportionnelle" esquissé ci-avant. Une analyse plus fouillée de ce modèle mène d'ailleurs au constat qu'il est, indépendamment de la procédure concrète de sélection, pas toujours réalisable (voir par exemple les réflexions émises dans Hartigan & Wigdor, 1989).

3 Le lecteur trouvera un tableau de ces comparaisons par exemple dans Schuler & Funke (1989) ou dans Smith & George (1992).

zinische Studiengänge" (TMS), remplit cette condition relative à ce problème de sélection. C'est dès lors, vu la situation décrite, la procédure la meilleure pour une sélection utile et équitable des candidats aux études de médecine en Suisse, procédure nécessitant bien entendu d'autres mesures de contrôle.[4]

La discussion en cours actuellement en Suisse montre toutefois que la conception qui vient d'être formulée est controversée dans de larges milieux. Dans un canton, un stage hospitalier a même été envisagé par votation populaire comme alternative au test, une procédure de sélection qui, cela est certain, ne résisterait pas à une analyse de l'utilité et de l'équité au sens indiqué plus haut. L'acceptabilité de procédures de sélection semble ainsi dépendre d'autres critères que de ceux utilisés jusqu'à présent. Il n'y encore que peu d'enquêtes empiriques concernant cette acceptabilité. Elles autorisent cependant quelques conclusions.

Fruhner, Schuler, Funke & Moser (1991) ont interrogé environ 1000 étudiants pour constater leurs préférences pour huit différentes procédures de sélection (de personnel). La méthode qui est apparue clairement la meilleure a été l'interview (entretien de présentation). Les tests psychologiques ont obtenu une note nettement moins bonne, mais meilleure que la procédure par tirage au sort, qui s'est classée en fin de liste (de même que les analyses graphologiques). Il est intéressant d'examiner ce qui se cache derrière ces préfé-

---

4 Il est peut-être indiqué d'ajouter une précision ici: le test mentionné suscite souvent la critique qu'il n'est pas axé sur le succès professionnel, mais sur le succès obtenu durant les études. Dans la perspective susmentionnée cette concentration sur le succès durant les études est cependant la seule voie praticable et psychologiquement raisonnable: toute procédure de sélection qui tente de se baser sur des critères de comportement dans la profession aurait sans aucun doute une valeur nettement moindre. Les raisons à l'appui de cette assertion sont entre autres les suivantes: 1.) Dans le cas d'études préparant à des activités professionnelles très diverses, il est extrêmement difficile de trouver des critères valables de manière égale pour les futures activités 2.) Il faudrait en outre chercher ces critères dans le *comportement du médecin*, c'est-à-dire non dans le domaine purement cognitif, mais dans le domaine dit de la personnalité, dont une appréhension valable est moins possible que dans une situation de sélection. 3.) De manière générale, il est admis, en outre, que la valeur baisse en fonction du temps qui s'écoule entre la sélection et les faits à prédire, vu que le développement humain peut très difficilement être prédit pour des périodes éloignées. Ces arguments indiquent qu'une procédure d'admission qui serait axée sur des critères du comportememt personnel lors d'activités professionnelles d'un futur lointain ne pourrait éviter de nombreuses décisions fausses, avec toutes les conséquences négatives que cela impliquerait pour l'utilité et l'équité de la procédure. Il est certainement justifié d'exiger des critères qui prennent en compte les activités professionnelles, mais en l'occurrence ce point-là ne devrait pas être lié au problème de l'admission à la formation. Il devrait être lié à la formation elle-même et à la sélection effectuée tout au long de la formation.

rences (à noter que la comparaison n'a porté que sur l'entretien et les tests). De manière générale, l'entretien a été jugé nettement plus positif et transparent que le test et un peu moins angoissant pour les candidats. Certaines des questions posées à des personnes expérimentées en la matière ont montré en outre d'autres différences en faveur de l'entretien: on est d'avis qu'il est mieux à même d'appréhender les capacités qui importent dans ce contexte et qu'il offre de meilleures possibilités d'influencer soi-même activement son propre résultat dans la procédure. De même, les candidats estiment qu'ils réussissent en moyenne mieux dans un entretien que dans un test.

L'acceptabilité d'une procédure de sélection semble ainsi dépendre, entre autres choses, surtout de deux critères. L'un concerne la transparence ou l'impression subjective de la valeur de la procédure. En psychologie, c'est une notion qui, au fil du temps, a reçu la désignation de "Augenscheinvalidität" ou valeur apparente, désignation provenant du fait que cette impression ne résiste généralement pas à un examen empirique rigoureux. Ainsi, l'interview menée librement et ouvertement (et il semble que les personnes interrogées se soient fondées sur cette base) doit être rangé vu, le résultat obtenu, parmi les procédures peu valables du point de vue de la qualité des prévisions, c'est-à-dire parmi les procédures engendrant relativement beaucoup de jugements erronés. L'autre critère concerne l'impression de pouvoir prendre influence soi-même activement et de manière ciblée sur le résultat de la procédure de sélection. L'élément en jeu ici est semble-t-il l'idée (exacte ou illusoire) que le candidat peut contrôler l'issue de la procédure. Qu'il s'agit bien de ce critère ressort aussi d'une étude de Latham & Finnegan (1993), qui se réfère à l'acceptabilité diverse (appelée "faisabilité" dans ladite étude) d'interviews non structurées et d'interviews structurées. On parle d'interviews structurées quand ces dernières, à la différence d'interviews non structurées, sont largement standardisées et qu'elles font suite à un catalogue donné de questions. Cette méthode mène à une valeur de prédiction clairement supérieure, mais aussi à une plus grande parenté avec un test. Dans cette enquête comme dans l'autre, les personnes interrogées ont accordé la préférence à la procédure aménagée plus librement, en arguant par exemple qu'un tel entretien leur permet de mieux montrer leur motivation (op. cit., p. 52). Les personnes interrogées ont parfaitement conscience du fait que la possibilité (admise) que le candidat peut exercer un meilleur contrôle du résultat de l'interview entraîne également des problèmes pour ce qui est de la valeur et de l'équité de cette procédure. La question suivante leur a été posée: "Si vous vouliez recourir contre la décision des responsables de la sélection, vos chances de succès seraient-elles plus grandes dans le cas d'une interview structurée ou dans celui d'une interview non structurée?" Elles ont mentionné beaucoup plus fréquemment l'interview non structurée.

Ces constats mènent à la conclusion que les procédures de sélection sont implicitement liées, semble-t-il, à un clair conflit d'intérêts. Sous l'angle de la

technique de sélection, il y a lieu, aussi dans l'intérêt des personnes concernées, d'accorder la préférence à une procédure de valeur prédictive sûre et prouvée, qui engendre aussi peu que possible de décisions fausses et qui puisse être objectivement qualifiée d'utile et d'équitable. D'autre part, on ressent le besoin d'avoir une procédure transparente, dont on ait l'impression, à l'usage, qu'elle soit valable et influençable dans son propre intérêt. Ces deux conceptions sont justifiées, d'une certaine manière. Le conflit semble néanmoins n'être pas facile à résoudre.

Littérature citée

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.

Le "test des tests" - résultats d'un essai effectué avec le test d'aptitudes en Suisse en langues allemande et française

Rainer Hofer, Daniel Ruefli & Klaus-D. Hänsgen

Centre pour le développement de tests et le diagnostic (CTD),

Université de Fribourg

Introduction

Le test d'aptitudes importé d'Allemagne et adapté à la Suisse pour servir de critère d'admission aux études de médecine en Suisse est, considéré dans son ensemble, une solution tant équitable que faisable (cf. les exposés de Klieme et de Trost). Un test psychologique, et le test d'aptitudes en est un, est assujetti à une vérification scientifique définie de manière précise. Il ne doit dès lors en aucun cas être admis ou toléré que le test désavantage des personnes testées. Une vérification du test en Suisse s'imposait parce que dans les discussions politiques relatives au numerus clausus ce n'était pas le numerus clausus en soi qui a été jugé inéquitable, mais le test destiné à servir de critère d'admission. Si le test était véritablement inéquitable, il devrait soit être modifié, soit être écarté comme critère d'admission.

L'essai effectué en Suisse

Le test effectué en Suisse (Hofer et al. 1995) visait dès lors en premier lieu à éclaircir les questions suivantes:

La **sélectivité en fonction de la performance** est-elle vraiment un critère d'admission adéquat? Ou le test est-il de manière générale trop facile ou trop difficile, si bien qu'il n'y a pas de possibilité de différencier sur la base des prestations au test?

Les **formes linguistiques** des tests sont-elles équivalentes et garantissent-elles la même fidélité et la même équité pour tous les groupes linguistiques?

Les performances réalisées lors des tests mettent-elles en évidence des différences entre les **sexes** ou les différents **types de maturité** telles qu'elles excluent la possibilité d'une égalité des chances?

L'essai a eu lieu au Collège Sainte-Croix à Fribourg. Il a été effectué avec des gymnasiens et des gymnasiennes germanophones et francophones se trouvant à une année des examens de maturité et qui s'étaient déclarés disposés à servir de cobayes. Les élèves ont été informés du déroulement du test une semaine avant

le jour du test. A cette occasion les différents types d'items ont été discutés. Il leur a été distribué également un Info-Test (Centre pour le développement de tests et le diagnostic 1995) pour leur donner la possibilité de se préparer au test à la maison. S'il est important de se familiariser avec les exigences requises au test, c'est en premier lieu pour éviter que le jour du test, les candidats ne perdent un temps considérable à comprendre les instructions et qu'ils puissent commencer immédiatement à traiter les items.

A la différence de la procédure allemande, cet essai a été effectué durant une demi-journée seulement, raison pour laquelle seuls sept des neuf sous-tests ont été donnés à résoudre. L'échantillon germanophone a dû résoudre des items d'un test original publié en version allemande (Institut für Test- und Begabungsforschung 1990) sous le titre Test für medizinische Studiengänge (TMS). Les items de la version traduite en français (Centre pour le développement de tests et le diagnostic 1996) ont été donnés à résoudre à l'échantillon francophone.

42 gymnasiens et gymnasiennes germanophones et 125 gymnasiens et gymnasiennes francophones ont traité le test avec ses six sous-tests (le test visant à établir les capacités de concentration n'a pas pu être évalué). La part des femmes a été respectivement de 45 et de 56 pour cent. Les personnes testées se sont vu attribuer un point pour chaque item coté effectué correctement. On pouvait de ce fait obtenir au maximum 118 points. Le groupe germanophone a obtenu en moyenne 55,4 points, l'écart-type étant de 11,5 points. Les données correspondantes obtenues par le groupe francophone ont été les suivantes: 57,3 points pour la valeur moyenne et 12,4 points pour l'écart-type.

Le test satistait aux critères de qualité

La précision des mesures du test et de ses sous-tests a été déterminée au moyen du "coefficent de CRONBACH". Il est apparu que dans le groupe germanophone les valeurs se situent aux alentours de la médiane de 0,71. Le groupe francophone présente pour le test global une médiane de 0,65. Cette valeur inférieure a été due en partie au sous-test ”Mémorisation de faits”, dont la valeur a été de 0,46. On a émis l'hypothèse que dans ce sous-test les aides mnémotechniques présentes dans la version germanophone pour la mémorisation des faits (Hänsgen et al. 1995) se sont perdues dans la traduction.

La comparaison des résultats de l'essai de test avec ceux obtenus par les candidats allemands qui avaient effectué le test en 1992, pris comme "groupe témoin" (Trost et al. 1990) avec une médiane, pour le test global, de 0,72, révèle que lors de cette épreuve de test une précision de mesure presque égale a été atteinte. Dans deux des six sous-tests, les échantillons suisses ont même obtenu en moyenne un résultat un peu meilleur que le groupe témoin allemand. Une analyse des données du test semble indiquer que les différences de niveau sont

probablement dues surtout au manque de motivation des participants à l'épreuve d'essai. Cependant, dans l'ensemble, la fiabilité correspond dans les deux groupes linguistiques aux exigences à satisfaire par une procédure psychodiagnostique. Mais, surtout, la fiabilité de la valeur globale du test utilisé pour l'admission correspond dans les deux groupes linguistiques, pour ce qui est du niveau, à celle du groupe témoin allemand.

Le test permet une différenciation optimale

Les répartitions des valeurs brutes, les degrés de difficulté et les possibilités de sélectivité permettent la différenciation des candidates et des candidats selon leur performance au test. En comparaison du groupe témoin de la République fédérale d'Allemagne (55%), un degré de difficulté de 48% a été atteint en moyenne. Vu la situation telle qu'elle se présente en Suisse, il faudrait une différenciation qui fasse en sorte que 75 à 85 pour cent des meilleurs au test puissent être admis aux études. Le test devrait dès lors différencier suffisamment bien dans ce secteur. Dans cette zone, au maximum 2,3 pour cent des personnes ont atteint la même valeur de points. Il serait dès lors possible de trouver avec une précision suffisante une valeur de points limite que seul excède un taux fixé selon la capacité.

Le test ne mesure pas la même chose que les notes scolaires

A partir des notes scolaires, il a aussi été possible de calculer dans quelle mesure les prestations au test et les prestations scolaires correspondent les unes aux autres. La corrélation relativement basse (0,23) de la valeur de test avec une note globale (demandée) montre que le test ne mesure pas exactement la même chose que ce qui est mis en lumière par les notes scolaires. Des corrélations du même genre obtenues en Allemagne sont un peu plus élevées (0,39 - cf. Trost et al. 1994). Cette différence entre les notes scolaires et le résultat du test, Trost et al. l'explique par des influences qui réduisent la corrélation et qui proviennent de la diversité des échelles d'appréciation valables pour les notes scolaires. En premier lieu le problème de la sévérité hétérogène d'appréciation observable en Suisse de canton à canton et d'école à école pour l'octroi de la note de maturité pourrait réduire encore la corrélation. Au contraire, les échelles d'appréciation du test sont uniformes et il serait possible de compenser ces différences. A ce sujet il faut tenir compte du fait que la valeur prédictive du test n'est en aucun cas inférieure à celle des notes scolaires pour la prévision du succès aux études.

Pour le reste, le contrôle des rapports entre les sous-tests montre que chaque sous-test teste des critères qui lui sont propres et que ce de fait aucun sous-test n'est superflu, parce que redondant.

A l'analyse des rapports entre le test et d'autres critères, le rapport entre type de maturité et résultat au test s'est révélé être significatif. Dans le groupe francophone, les personnes à maturité de type D ont obtenu un résultat moyen considérablement inférieur à celui des personnes qui ont achevé leur gymnase avec une maturité des types A à C. Cela correspondrait à des résultats comparables pour les études de médecine: par exemple, à la Faculté de médecine de l'Université de Berne, un taux variant entre 44% et 50% des titulaires d'un certificat de maturité des types A à C n'a pas réussi le premier examen propédeutique du premier coup, alors que chez les titulaires d'un certificat du type D, 77% des candidats ont réussi à la première tentative (Hofer 1992).

Le test ne défavorise pas les femmes

Il est exigé à juste titre que les hommes et les femmes aient les mêmes chances d'admission aux études universitaires et que les progrès obtenus ne soient pas compromis par l'application du test. Dans certains exposés a été exprimé le soupçon que le test discrimine les femmes. Ces jugements portés sur le test ont cependant toujours fait état des résultats du test allemand, qui n'ont toutefois pas pu être confirmés ici.

Les résultats de l'épreuve d'essai n'indiquent pas de différences systématiques pour les deux sexes. Il n'a pu être prouvé ni au niveau du test, ni à celui des sous-tests, ni à celui des items qu'il y a lieu d'admettre, le cas échéant, que les femmes subissent des désavantages par rapport aux hommes en Suisse. L'analyse des items selon Mantel & Haenszel (1959) n'a présenté une distorsion liée au sexe que pour 4 des 136 items dans le groupe germanophone et pour 3 des 136 items du groupe francophone. Pour les deux sexes, le contenu, la séléctivité, le modèle de réponse et le degré de difficulté des différents items ont été contrôlés. On n'a cependant pas trouvé d'interpréations concluantes pour ces distorsions d'items. Il doit dès lors s'agir d'un hasard statistique.

Au niveau des sous-tests, les femmes du groupe germanophone ont obtenu en moyenne, dans 5 et 6 des 6 sous-tests, un résultat meilleur que les hommes dans leur groupe linguistique (dont 3 sont significatifs du point de vue statistique, $p < 0,05$, tableau 1). Dans le groupe francophone, la comparaison avec le groupe témoin au niveau des sous-tests est équilibrée. Alors que les femmes ont obtenu un meilleur résultat dans les sous-tests "Compréhension de textes", "Mémorisation de figures" et "Mémorisation de faits", la situation a été inverse dans les sous-tests "Reconnaissance de fragments de figures", Figures tubulaires" et "Problèmes quantitatifs et formels". Au niveau du test, les femmes du groupe germanophone et les hommes du groupe francophone ont atteint en moyenne de meilleurs résultats que les personnes du même sexe. En regard du groupe témoin de RFA (Trost et al. 1994; Hofer et al. 1995) on

constate que le groupe germanophone se comporte exactement de manière inverse.

Tableau 1: Rapport entre respectivement sexe et langue d'une part et résultat du test au niveau du test et à celui des sous-tests d'autre part

	Groupe germanophone					**Groupe francophone**				
Sous-test / Test	Hommes (n=23)		Femmes (n=19)			Hommes (n=48)		Femmes (n=60)		
	m	**s**	**m**	**s**	**p**	**m**	**s**	**m**	**s**	**p**
Compréhension de textes	9,04	3,91	10,37	3,15		9,60	3,04	9,80	2,83	
Mémorisation de figures	6,87	2,67	8,84	2,39	< 0,05	8,04	3,61	8,90	3,24	
Mémorisation de faits	7,26	2,86	10,26	3,83	< 0,05	10,52	2,79	11,07	2,99	
Reconnaissance de fragments de figures	11,35	3,52	13,16	3,24	< 0,05	12,17	3,71	11,73	2,83	
Figures tubulaires	9,09	3,58	9,11	3,57		9,92	3,94	7,83	3,43	< 0,05
Problèmes quant. et formels	8,26	3,98	7,84	4,36		9,29	4,26	6,87	3,00	< 0,05
Test dans son ensemble	51,87	11,52	59,58	12,01	< 0,05	59,54	14,55	56,20	10,84	

L'égalité des chances vaut aussi pour les groupes linguistiques

Le test original a été mis au point en Allemagne. En reprenant un tel test dans une autre culture et une autre langue, il est possible que l'on crée des différences qui influent sur l'égalité des chances. Il est donc nécessaire que deux groupes d'experts indépendants surveillent la traduction et la rétroversion des items de manière à éviter des pertes ou des déformations de leur contenu. Ce n'est qu'après l'exécution du test que l'on peut vérifier, à l'examen des paramètres de contrôle quantitatifs, si l'adaptation a entraîné des défauts.

Il n'y a que peu de différences spécifiquement linguistiques (tableau 1) entre les deux groupes suisses. Dans les sous-tests "Mémorisation de figures" et "Mémorisation de faits", le groupe francophone a obtenu un résultat meilleur que le groupe germanophone du Collège Sainte-Croix. On notera que la différence observée dans le sous-test "Mémorisation de faits" est significative. Dans les autres sous-tests il n'y avait pas de différence entre les résultats des deux sexes.

Dans l'analyse selon Mantel & Haenszel (1959), 6 des 136 items ont présenté une distorsion relative à la langue, dont trois proviennent de sous-tests plutôt indépendants de la langue. La vérification de ces sous-tests montre que les différences ne sont pas véritablement systématiques. Dans les sous-tests

"Compréhension de textes", "Reconnaissance de fragments de figures", "Figures tubulaires" et "Problèmes quantitatifs et formels", le groupe germanophone a obtenu un résultat légèrement meilleur que le groupe francophone. Ce dernier a toutefois réalisé en moyenne. dans les sous-tests de mémoire "Mémorisation de figures" et "Mémorisation de faits", un nombre de points supérieur, qui lui permet d'enregistrer un avantage sur l'ensemble du test par rapport à l'autre groupe linguistique. La traduction soignée et coûteuse du test en langue française a donc donné un résultat qui permet de parler d'une adaptation réussie dans une autre culture.

Conclusion

l'épreuve a permis de confirmer qu'il est possible de satisfaire en Suisse aux conditions structurelles et techniques que requiert ce test et qu'il s'agit d'une procédure équitable.

1. En se basant sur les valeurs globales du test, on obtient un critère d'admission aux études de médecine qui présente une sélectivité suffisante. La difficulté du test pour les deux groupes linguistiques suisses diffère de peu de celle rencontrée par le groupe témoin allemand.
2. L'adaptation du test pour la Suisse alémanique et pour la Suisse romande atteint, sur le plan de la fiabilité et de l'équité, une qualité comparable à la version originale allemande. Il a pu ainsi être prouvé également que la procédure d'adaptation choisie est adéquate.
3. L'épreuve n'a pas permis de confirmer des différences de valeurs du test pour ce qui concerne le sexe. Il apparaît d'entrée de cause que les chances sont égales pour les hommes et les femmes.

Littérature

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 médecine. 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.