Medical Education 1989, **23**, 147–151

Reliability of short-listing in medical student selection

I. C. McMANUS & P. RICHARDS

St Mary's Hospital Medical School, Imperial College of Science, Technology and Medicine, London

Summary. One in eight Universities Central Council on Admissions (UCCA) applications for admission to St Mary's Hospital Medical School in 1986 were in due course recirculated to the four short-listers, being seen again either by the same short-lister or by another short-lister. Intrarater reliabilities were high, not only for measures of educational achievement, but also for the more subjective assessments. Interrater reliabilities were more variable, being very high for educational achievement, but rather lower for the more subjective items, and being lowest for the assessment of 'Interests', suggesting divergence between short-listers' perceptions of the terms. Nevertheless, all reliabilities were sufficiently high to justify the continued use of these criteria during selection.

Key words: *school admission criteria; *judgement; faculty, medical; London

Introduction

Applicants to study medicine in British medical schools apply through the Universities Central Council on Admissions (UCCA), submitting a standard three-page application form (the 'UCCA form') which is photocopied by UCCA and distributed to the five universities or medical schools chosen by the applicant. On the basis of this form applicants are either short-listed for interview (in about 70% of medical schools), or in non-interviewing schools are made offers directly, either unconditionally, or conditional upon examination performance.

Correspondence: Dr I. C. McManus, Academic Department of Psychiatry, St Mary's Hospital Medical School, Paterson Wing, Praed Street, London W2 1NY, UK.

Despite the prime importance of short-listing in selection, there have been few studies of the limited information available in the UCCA form, or of the reliability of judgements made from it. Short-listers at St Mary's have, for a number of years, used a *pro forma* on which they used 5-point scales to rate eight separate items concerning the candidate: O- and A-level grades, 'Interests', 'Contribution to school', 'Achievement', 'Contribution to community', 'Referee's report' and 'Potential'. Although eight separate items are assessed, that does not ensure that eight statistically independent types of information are extracted about each candidate; for instance it might be that a short-lister rates the O-levels, and according to whether they are good or bad then rates that candidate as good or bad on each of the other seven items, so that a single dimension, scale or factor underlies the eight observations. The number of statistically independent factors, or dimensions, can be determined by factor analysis. In a previous study (McManus & Richards 1984b) we have shown that one short-lister (PR) using an eight-item *pro forma* extracted three statistically independent types of information, about 'Academic ability', 'Interests', and 'Contribution to community', each of which contributed separately to the decision about whether or not the applicant should be interviewed. In a second study (McManus *et al.* 1989) we showed that three other short-listers also used the same essential factors, and the PR's judgements had the same factorial structure 5 years later.

Judgements can be reliable in two separate senses: agreement of the assessor with him- or

herself ('intrarater') and agreement with others ('interrater'). It is possible for intrarater reliability to be high, and interrater reliability to be low, in which case an individual's judgements are consistent but idiosyncratic. The issue of reliability (i.e. of consistency between judgements) should not be confused with the separate notion of validity (the degree to which a judge's assessment of the candidate's characteristic corresponds with the true or actual characteristic of the candidate [Ghiselli *et al.* 1981]). Here we only consider reliability, which is a necessary precondition of validity (although it is not sufficient to ensure validity).

We report here a study of short-listers' judgements of UCCA forms which was designed to allow assessment of reliability within and between short-listers. Elsewhere we have described differences between short-listers in their overall distribution of assessments, and in the factor structure of their assessments (McManus *et al.* 1989).

Methods

In the autumn of 1985 we carried out a survey of the selection of students applying for admission to St Mary's in October 1986. The survey was broadly similar in structure to that carried out in 1980 of applicants for admission in 1981 (McManus & Richards 1984a,b). An important change was that partly because of increased numbers of applicants, short-listing was carried out by four individuals rather than one. One of the short-listers (A) was the Dean (PR), who had previously carried out the short-listing single handed. Another short-lister (B) had experience of short-listing from previous years, while the remaining two (C, D) were carrying it out for the first time in 1985.

Immediately after receipt at St Mary's, UCCA forms were allocated randomly, in a predetermined sequence, to short-listers who completed a *pro forma* on the application, and made a decision about interviewing. For approximately one in eight of applicants the form was assessed again by a short-lister, in half of the cases by the same short-lister (in which case the form was held to one side for 3 or 4 weeks before reassessment, to reduce memory effects), and in half of the cases by one of the other three short-listers. Short-listers were unaware which forms were to be recirculated, or whether a form had previously been circulated, and no evidence as to the first judgement was present at the second circulation. In order that the system should be fair to all candidates, decisions about interviews were based entirely on the first short-lister's judgement.

Applicants were only included in the overall study of selection if they had British postal addresses for correspondence, and were only included in this study of short-listing if they were also UK nationals (in order to reduce memory effects for applicants who might be relatively unusual). A weighted system of allocation was used so that forms from earlier applicants were somewhat more likely to be recirculated than later applicants. This was done because there is an excess of better qualified applicants within those applying early (McManus & Richards 1984a).

Short-listers made their judgements on a one-page *pro forma* which asked them to assess each of eight separate characteristics on a 5-point scale (Excellent — top 10%; Good — top 10–30%; Adequate — middle 30–70%; Poor — bottom 70–90%; and Bad — below 90% of applicants). A-level achievement was scored both for examinations already taken, and for predictions of those still to be taken, and these two measures were also combined into a single composite measure. The overall decision of the short-lister was on a 3-point scale: (A) Definitely interview (20% of candidates); (B) Possibly interview (10% of candidates); and (C) Reject (70% of candidates).

Statistical analysis

The relationships between the judgements can be represented by a path diagram (Fig. 1) (Kenny 1979). Path diagrams allow correlations between measured or unmeasured variables to be read off directly by following along all possible routes between the two variables; thus the correlation between 'judge A's perception' and 'judge B's perception' in Fig. 1 is $b \times b$, and the correlation between 'judge A rating 1' and 'judge B rating 1' is $a \times b \times b \times a$. By making the structural relationships between variables explicit the path diagram

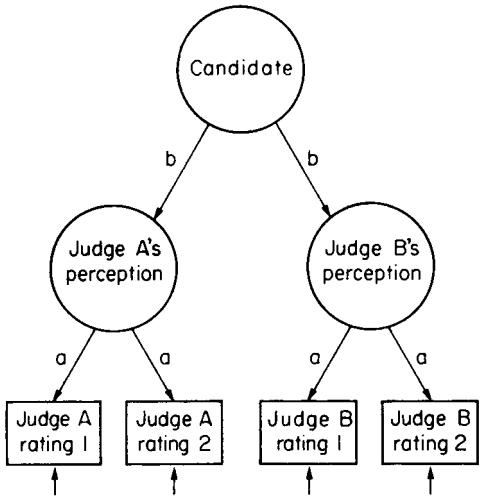

Figure 1. Shows the formal structural model underlying separate assessments of a single candidate by two judges on separate occasions. The ratings of each individual judge (the measured variables shown in the rectangles) are determined by the judge's perceptions (latent, unmeasurable variables), which are themselves determined by the candidate's true characteristic (also a latent variable). Each rating also has a random, unsystematic component, indicated by the short vertical arrows.

helps to visualize and clarify the often obscure covariance algebra which is implicit in any analysis of reliability.

Judge A assesses a candidate on two occasions, as does judge B. The two ratings made by judge A are each imperfect manifestations of A's perception of the candidate, as also are those of B. A and B's perceptions are themselves imperfect manifestations of the candidate him- or herself. The path from each judge's perception to his judgement, a, is a measure of intrarater reliability, and the path from the candidate to the judge's perception, b, is a measure of interrater reliability. Let r_w be the within-rater correlation of judgements, and r_b the between-rater correlation of judgements. From the path diagram, $r_w = a \times a$, and hence $a = \text{sqrt}(r_w)$. Similarly $r_b = a \times a \times b \times b$, and therefore $b = \text{sqrt}(r_b)/a = \text{sqrt}(r_b/r_w)$. The measures of reliability can be seen as the regressions of separate, repeated measures upon the true or latent variable.

Results

2210 applicants who had UK postal addresses and applied before 15 December 1985 were included in the selection survey. 273 UCCA forms were seen by two short-listers, 136 by the same person who had previously short-listed them, and 137 by a different short-lister.

Table 1 shows the correlations between the separate judgements made by a single short-lister and those made by different short-listers. From these values the intra- and interrater reliabilities were calculated as described above.

The analyses of Table 1 assume that the reliabilities of the individual short-listers were all similar. That assumption was tested by computing separately the correlations between the two judgements made by each of the short-listers, and comparing each correlation with the sampling distribution expected on the basis of the overall correlations reported in Table 1: only 1 out of 36 correlations was significantly different from the group value at the 5% level. A similar analysis was carried out for the judgements made by two different short-listers, correlations being calculated separately for all the possible six pairs of short-listers. Of 54 correlations only five were significantly different from the value obtained by combining across all short-listers. Thus of 90 individual correlations, only 6 (6·6%) were significantly different at the 5% level from values based on combined data. There is therefore no evidence that these short-listers differ from one another in their intra- or interrater reliabilities.

Discussion

Short-listing of applicants on the strength of UCCA forms is an unavoidable part of a selection system in which there are many more applicants than can either be accepted or interviewed, and in which entry is not to be determined only by academic examination ranking. The making of broader judgements of the suitability of candidates for entry to a caring profession is necessarily an imprecise task under any circumstances and is particularly difficult from the limited description given in a written document. Nevertheless, we have found that the judgements have a reasonable degree of reliability, both between and within short-listers.

Table 1. Correlations between separate assessments of UCCA forms either by the same short-lister or by different short-listers, and calculated intrarater and interrater reliabilities (see text)

	Correlations		Reliability	
Scale	Same short-lister	Different short-lister	Intrarater	Interrater
O-levels	0·839	0·823	0·916	0·991
A-levels taken	0·831	0·916	0·911	1·050†
A-level predictions	0·827	0·862	0·910	1·021†
A-levels overall	0·851	0·878	0·922	1·016†
Interests	0·665	0·225	0·815	0·581
Contribution to life of school	0·627	0·367	0·792	0·765
Achievement	0·677	0·463	0·823	0·828
Contribution to community	0·664	0·402	0·815	0·778
Referee's report	0·676	0·316	0·822	0·684
Potential	0·684	0·408	0·827	0·773
Final decision	0·663	0·403	0·814	0·785

† Indicates theoretically inadmissable values greater than unity, occurring due to sampling variation, and which should be interpreted as being unity.

It is not possible to state any absolute minimum above which a reliability should lie, for as long as the value is greater than zero then *some* reliable information is being extracted. Nevertheless, as a rough guide it can be noted that if two judgements share 50% of their variance, they have a correlation of 0·707 and hence a reliability of 0·841. It would be desirable if intrarater reliabilities were at least at this level. Similarly interrater reliabilities should be at least as reliable as intrarater reliabilities, so that the variance between raters is no greater than that within a single rater.

Even the least reliable of the intrarater assessments, 'Contribution to school', shares 40% of variance between separate assessments by the same individual. However interrater reliabilities were rather more variable, varying from values around unity for academic achievement (i.e. no disagreement between short-listers in their criteria), down to a value of 0·581 for 'Interests', implying only 11% of shared variance between short-listers in their interpretation of this term. That interrater reliabilities of non-academic terms are less than intrarater reliabilities suggests the need for more explicit definition of the meaning of terms, perhaps with examples and training in their use. Experience *per se* does not improve interrater reliability since correlations between the two most experienced short-listers were no higher than those between the least. Similarly intrarater reliabilities did not differ according to experience, suggesting that much of the variability is intrinsic to the task itself.

Assessments based on informal judgements are rarely very precise measures (for instance never achieving the typical reliabilities reported for psychometric tests of 0·95 and higher). Nevertheless, in the absence of such psychometric tests for those broad personality characteristics which both medical selectors and applicants consider to be important in selection for medical practice, it would appear that there is sufficient agreement between short-listers to justify the continued use of assessments. Improvements in interrater reliability could almost certainly be achieved by more explicitly defined criteria for the specific scales, perhaps using 'behaviourally anchored rating scales' (Schwab *et al.* 1975).

Acknowledgements

We are grateful to our short-listers, Dr A. W. Boylston, Dr B. P. Curwain and Dr G. H. Tait for so readily agreeing to the extra workload imposed by the additional assessments required

by this study. We also thank Mrs R. Boyd and Ms C. Richards for their help in the administration of the study and the Economic and Social Research Council for financial support.

References

Ghiselli E.E., Campbell J.P. & Zedeck S. (1981) *Measurement Theory for the Behavioural Sciences.* W.H. Freeman, San Francisco.

Kenny D.A. (1979) *Correlation and Causality.* John Wiley, New York.

McManus I.C., Maitlis S.L. & Richards P. (1989) Short-listing of applicants from UCCA forms: the structure of pre-selection judgements. *Medical Education* **23,** 136–46.

McManus I.C. & Richards P. (1984a) Audit of admission to medical school: I. Acceptances and rejects. *British Medical Journal* **289,** 1201–4.

McManus I.C. & Richards P. (1984b) Audit of admission to medical school: II. Short-listing and interviews. *British Medical Journal* **289,** 1288–90.

Schwab D.P., Heneman H.G. & De Cotiis T. (1975) Behaviourally anchored rating scales: a review of the literature. *Personnel Psychology* **28,** 549–62.

Received 17 *March* 1988; *editorial comments to authors* 21 *April* 1988; *accepted for publication* 15 *July* 1988