Medical Education 1989, **23**, 136–146

Short-listing of applicants from UCCA forms: the structure of pre-selection judgements

I. C. McMANUS, S. L. MAITLIS & P. RICHARDS

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

Summary. Applicants for admission to St Mary's Hospital Medical School in 1986 were short-listed for interview by one of four assessors, who each made their assessments on a nine-item *pro forma*. One short-lister had also been studied in detail during 1981. Short-listers used the full range of possible judgements, in approximately the proportions requested. Only minor differences were found between them in the mean and range of their judgements, suggesting that similarity of standards can be maintained while using a number of separate short-listers. A confirmatory factor analysis of individual short-listers' judgements showed that all were extracting three separate factors, named 'Academic ability', 'Interests' and 'Contribution to community', although the less experienced short-listers differentiated these items less well than the more experienced. The short-lister assessed in 1981 and 1986 had retained an almost identical factor structure over the 5-year period.

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

Introduction

Many more people apply for admission to medical schools in the United Kingdom than can be accepted. In current methods of student selection, short-listing, or pre-selection (Herriot 1982), plays a central role, both in schools which do not interview (where it is the only hurdle to be jumped, apart from achievement of high examination grades), and in those which do interview, since only a minority of applicants are short-listed, and the majority of those short-listed are eventually accepted (see for example McManus & Richards **1984a**). That short-listing is numerically important does not of course guarantee its validity as a means of selecting medical students; that can only be ascertained by a systematic trial against some other method such as random selection, and to our knowledge that has never been carried out.

Despite the importance of short-listing, there have been few studies of its reliability or validity, using the terms in the technical senses usually ascribed to them in psychometrics (e.g. Ghiselli *et al.* 1981, pp. 191 & 266) of reliability as the degree to which individual short-listers agree on which candidates are good or bad, and validity as the relation of the assessments to other measures of the candidates' present or future behaviour. Neither have there been assessments of the types or the amount of information on which short-listing decisions are based.

Short-listing of applicants for medical school admission has been analysed in some detail in the 1981 cohort of applicants to St Mary's Hospital Medical School in the University of London (McManus & Richards 1984b). In that study one person (PR, Dean of the Medical School) short-listed all applicants using an eight-item *pro forma*, each candidate being rated on a 5-point scale for each of the eight items. Conventional exploratory factor analysis followed by a Varimax rotation revealed the presence of three independent dimensions which we termed 'Academic

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

ability', 'Interests', and 'Contribution to community'. These three factors accounted for 78% of the total variance, and each also contributed independently to the decision as to whether or not a candidate should be interviewed. Subsequent analyses showed that these three measures did not predict success during preclinical examinations (McManus & Richards 1986a), although they could help to explain why applicants with non-European surnames fared less well overall in their applications (McManus & Richards 1985).

In reporting the results of our previous study of short-listing we were aware of several deficiencies in the design: the assessments were made by a single short-lister, and results might therefore be idiosyncratic; we had no measures of the reliability of the component assessments from which the three measures were derived; and since the factor analysis was entirely exploratory, rather than attempting to confirm a particular hypothesis, the statistical problem therefore arose that the measures obtained were necessarily orthogonal (statistically uncorrelated) as a consequence of the statistical method. We hope now to remedy several of these deficiencies, by reporting a second study of short-listing on a new and larger cohort of applicants (for admission in October 1986), who were short-listed by four short-listers rather than one. In this paper we consider whether short-listers agree in the overall levels of their assessments and how the factor structure of the assessments differs between individuals. Since one of the four short-listers was PR we can also assess the long-term stability of one short-lister's factorial structure over a 5-year period. In a separate paper we have considered the question of the reliability of the judgements made, a proportion of UCCA forms having been recirculated to the same or another short-lister to assess intra- and inter-short-lister reliabilities (McManus & Richards 1989).

Two logically and statistically separate types of question can be asked about the judgements made by short-listers: do they differ in their first-order statistics (i.e. means, variances, and other distributional characteristics); and do they differ in the interrelationships between the various judgements (i.e. second-order statistics, typically correlations). The first type of statistic will tell us whether some short-listers are 'doves' and others are 'hawks'. This is important in a system in which most application forms are assessed in detail by only a single short-lister, since if there are systematic differences between short-listers then some applicants might be treated unfairly. The second type of statistic answers a very different type of question concerning the structure and the dimensionality of each short-lister's judgements; short-listers may differ in that one may tend merely to rate all applicants as globally 'good' or 'bad', using a single underlying dimension to the judgements, whereas another may make several statistically independent assessments ('good on scale A but poor on scale B'), which are then combined into a final decision. The UCCA form is a complex document with a lot of explicit and implicit information on it, and therefore it is quite possible that short-listers will differ in the types and quantity of information that they extract from it.

Methods

During the autumn of 1985 St Mary's Hospital Medical School received 2399 applications for admission in October 1986. All applicants who had included a British postal address for correspondence were included in the survey of selection (n = 2211), and these applicants had their UCCA forms assessed by one of four short-listers who had been allocated at random, in a pre-arranged order. Not all UCCA forms were suitable for the present analysis for reasons such as incomplete information or unusual examination qualifications. Short-listers used a special *pro forma* to assess each applicant on nine separate scales, for each of which five possible responses were possible: 'Excellent: top 10%'; 'Good: 10th–30th percentile'; 'Adequate: 30th–70th percentile'; 'Poor: 70th–90th percentile'; or 'Bad: bottom 10%'. These nine scales were identical to those used previously, and described elsewhere (McManus & Richards 1984b), with the minor exception that the single item called 'A-level grades' in 1981 was divided in 1986 into two categories, 'A-levels taken' and 'A-level predictions'. The final item on the *pro forma* asked for a decision using one of the responses: 'Definite interview: 20% of candidates'; 'Possible inter-

view: 10% of candidates'; or 'Definitely do not interview: 70% of candidates'. These proportions were calculated to be compatible with the maximum number of candidates that it was practical to interview. Short-listers were also asked to indicate whether they had any personal knowledge of the applicant, their parents or their school.

The four short-listers differed in their experience of the task. Short-lister A (PR) was very experienced, having been the sole short-lister for admission from October 1979 until 1984. Short-lister B had short-listed for admission in October 1985, and short-listers C and D were short-listing for the first time. Each short-lister received about 600 application forms, and in addition each also received some forms on a second occasion which either they or other short-listers had seen previously. The present analysis does not consider these repeat assessments, which are the subject of another study (McManus & Richards 1989). Each short-lister saw about 660 UCCA forms during the period from October 1985 to March 1986, of which the majority were seen during October to December.

In order to ensure that a similar proportion of applicants was recommended for interview by each short-lister all completed *pro formas* were scanned by the Dean (PR, short-lister A) who made a 'Dean's decision' (on a 3-point scale) and then a 'final decision' (on a 2-point scale) for each candidate.

Statistical analysis

Conventional descriptive statistics and analysis of variance were carried out using the SPSS-X program package (Anon. 1983), and this was also used to generate correlation matrices for confirmatory factor analysis which was carried out using LISREL (Joreskog & Sorbom 1983).

Results

Analysis of means

Table 1 shows the various scales on which judgements were made, and the proportions of judgements made in each of the various categories in relation to the requested distributions. For most candidates an assessment was made only of 'A-level predictions' (if they were pre-A-level at the time of application) or of 'A-level results' (if they were post-A-level). The two scales were combined into a single scale 'A-levels'. A few resit candidates had assessments made of both scales, and the composite scale was then based on their actual results. It can be seen that broadly speaking the recommended proportions have been achieved, although short-listers found it difficult to rate candidates on non-academic scales as 'excellent' or 'bad', perhaps reflecting a lack of precise information on the form itself. The use of five categories had achieved its specific object of attaining a wide spread in judgements, so that sufficient variance was present to be able to discriminate between candidates.

If short-listers are 'hawks' then they will be expected to use the 'excellent' and 'good' categories less often than if they were 'doves'. This we assessed by scoring the five categories (5 = 'excellent'; 1 = 'bad') and calculating a mean score for each short-lister on each scale. Assuming that the quality of candidates randomly sent to each short-lister was truly equivalent, then differences in means should reflect differences between hawks and doves. Data were analysed by a one-way analysis of variance, with a Scheffé *post-hoc* comparison for differences in means of all possible pairs of short-listers. Table 2 summarizes these analyses, an arrow between two short-listers indicating the direction of a significant difference. For all scales there are significant differences present, and there is a consistent pattern whereby short-lister D is the most hawkish and short-lister B the most dovish. However, it must be emphasized that although statistically significant, these effects are very small, accounting in general for very small proportions of the total variance, the two largest accounting for only 10% of variance. The differences were particularly small for the overall decision, upon which the interview decision was based.

Short-listers might also differ in the variance of their judgements, so that some might use the categories of excellent and bad a lot of the time, whereas others might crowd most of their judgements into the middle categories. Table 2 also summarizes analyses of difference in variance between the four short-listers, expressed as the

ability', 'Interests', and 'Contribution to community'. These three factors accounted for 78% of the total variance, and each also contributed independently to the decision as to whether or not a candidate should be interviewed. Subsequent analyses showed that these three measures did not predict success during preclinical examinations (McManus & Richards 1986a), although they could help to explain why applicants with non-European surnames fared less well overall in their applications (McManus & Richards 1985).

In reporting the results of our previous study of short-listing we were aware of several deficiencies in the design: the assessments were made by a single short-lister, and results might therefore be idiosyncratic; we had no measures of the reliability of the component assessments from which the three measures were derived; and since the factor analysis was entirely exploratory, rather than attempting to confirm a particular hypothesis, the statistical problem therefore arose that the measures obtained were necessarily orthogonal (statistically uncorrelated) as a consequence of the statistical method. We hope now to remedy several of these deficiencies, by reporting a second study of short-listing on a new and larger cohort of applicants (for admission in October 1986), who were short-listed by four short-listers rather than one. In this paper we consider whether short-listers agree in the overall levels of their assessments and how the factor structure of the assessments differs between individuals. Since one of the four short-listers was PR we can also assess the long-term stability of one short-lister's factorial structure over a 5-year period. In a separate paper we have considered the question of the reliability of the judgements made, a proportion of UCCA forms having been recirculated to the same or another short-lister to assess intra- and inter-short-lister reliabilities (McManus & Richards 1989).

Two logically and statistically separate types of question can be asked about the judgements made by short-listers: do they differ in their first-order statistics (i.e. means, variances, and other distributional characteristics); and do they differ in the interrelationships between the various judgements (i.e. second-order statistics, typically correlations). The first type of statistic will tell us whether some short-listers are 'doves' and others are 'hawks'. This is important in a system in which most application forms are assessed in detail by only a single short-lister, since if there are systematic differences between short-listers then some applicants might be treated unfairly. The second type of statistic answers a very different type of question concerning the structure and the dimensionality of each short-lister's judgements; short-listers may differ in that one may tend merely to rate all applicants as globally 'good' or 'bad', using a single underlying dimension to the judgements, whereas another may make several statistically independent assessments ('good on scale A but poor on scale B'), which are then combined into a final decision. The UCCA form is a complex document with a lot of explicit and implicit information on it, and therefore it is quite possible that short-listers will differ in the types and quantity of information that they extract from it.

Methods

During the autumn of 1985 St Mary's Hospital Medical School received 2399 applications for admission in October 1986. All applicants who had included a British postal address for correspondence were included in the survey of selection (n = 2211), and these applicants had their UCCA forms assessed by one of four short-listers who had been allocated at random, in a pre-arranged order. Not all UCCA forms were suitable for the present analysis for reasons such as incomplete information or unusual examination qualifications. Short-listers used a special *pro forma* to assess each applicant on nine separate scales, for each of which five possible responses were possible: 'Excellent: top 10%'; 'Good: 10th–30th percentile'; 'Adequate: 30th–70th percentile'; 'Poor: 70th–90th percentile'; or 'Bad: bottom 10%'. These nine scales were identical to those used previously, and described elsewhere (McManus & Richards 1984b), with the minor exception that the single item called 'A-level grades' in 1981 was divided in 1986 into two categories, 'A-levels taken' and 'A-level predictions'. The final item on the *pro forma* asked for a decision using one of the responses: 'Definite interview: 20% of candidates'; 'Possible inter-

view: 10% of candidates'; or 'Definitely do not interview: 70% of candidates'. These proportions were calculated to be compatible with the maximum number of candidates that it was practical to interview. Short-listers were also asked to indicate whether they had any personal knowledge of the applicant, their parents or their school.

The four short-listers differed in their experience of the task. Short-lister A (PR) was very experienced, having been the sole short-lister for admission from October 1979 until 1984. Short-lister B had short-listed for admission in October 1985, and short-listers C and D were short-listing for the first time. Each short-lister received about 600 application forms, and in addition each also received some forms on a second occasion which either they or other short-listers had seen previously. The present analysis does not consider these repeat assessments, which are the subject of another study (McManus & Richards 1989). Each short-lister saw about 660 UCCA forms during the period from October 1985 to March 1986, of which the majority were seen during October to December.

In order to ensure that a similar proportion of applicants was recommended for interview by each short-lister all completed *pro formas* were scanned by the Dean (PR, short-lister A) who made a 'Dean's decision' (on a 3-point scale) and then a 'final decision' (on a 2-point scale) for each candidate.

Statistical analysis

Conventional descriptive statistics and analysis of variance were carried out using the SPSS-X program package (Anon. 1983), and this was also used to generate correlation matrices for confirmatory factor analysis which was carried out using LISREL (Joreskog & Sorbom 1983).

Results

Analysis of means

Table 1 shows the various scales on which judgements were made, and the proportions of judgements made in each of the various categories in relation to the requested distributions. For most candidates an assessment was made only of 'A-level predictions' (if they were pre-A-level at the time of application) or of 'A-level results' (if they were post-A-level). The two scales were combined into a single scale 'A-levels'. A few resit candidates had assessments made of both scales, and the composite scale was then based on their actual results. It can be seen that broadly speaking the recommended proportions have been achieved, although short-listers found it difficult to rate candidates on non-academic scales as 'excellent' or 'bad', perhaps reflecting a lack of precise information on the form itself. The use of five categories had achieved its specific object of attaining a wide spread in judgements, so that sufficient variance was present to be able to discriminate between candidates.

If short-listers are 'hawks' then they will be expected to use the 'excellent' and 'good' categories less often than if they were 'doves'. This we assessed by scoring the five categories (5 = 'excellent'; 1 = 'bad') and calculating a mean score for each short-lister on each scale. Assuming that the quality of candidates randomly sent to each short-lister was truly equivalent, then differences in means should reflect differences between hawks and doves. Data were analysed by a one-way analysis of variance, with a Scheffé *post-hoc* comparison for differences in means of all possible pairs of short-listers. Table 2 summarizes these analyses, an arrow between two short-listers indicating the direction of a significant difference. For all scales there are significant differences present, and there is a consistent pattern whereby short-lister D is the most hawkish and short-lister B the most dovish. However, it must be emphasized that although statistically significant, these effects are very small, accounting in general for very small proportions of the total variance, the two largest accounting for only 10% of variance. The differences were particularly small for the overall decision, upon which the interview decision was based.

Short-listers might also differ in the variance of their judgements, so that some might use the categories of excellent and bad a lot of the time, whereas others might crowd most of their judgements into the middle categories. Table 2 also summarizes analyses of difference in variance between the four short-listers, expressed as the

Table 1. Percentages of applicants being rated on each of the positions of the separate assessment scales in relation to the recommended proportions in each of the categories

Scale	Response categories and recommended proportions: Excellent (10%)	Good (20%)	Adequate (40%)	Poor (20%)	Bad (10%)	*n*
'O-levels'	9·5	30·9	41·6	16·1	2·0	2126
'A-level results'	7·1	14·7	24·8	33·6	19·8	464
'A-level predictions'	10·8	34·4	29·3	20·9	4·6	1656
'A-levels'	10·1	30·2	28·1	23·7	8·0	2084
'Interests'	2·4	26·6	56·8	13·6	0·7	2139
'Contribution to school'	1·4	26·0	53·3	18·2	1·1	2148
'Achievement'	2·6	18·4	57·9	20·2	0·9	2151
'Contribution to community'	1·6	23·5	48·8	22·1	4·0	2157
'Referee's report'	3·9	33·4	44·6	16·6	1·5	2163
'Potential'	1·0	23·6	45·2	27·9	2·3	2131
	Definite interview (20%)		Possible interview (10%)	No interview (70%)		
Short-listers' decision	20·1		12·3	67·5		2179
Dean's decision	11·1		23·1	65·7		2179
			Interview	No interview		
Final decision			20·9%	79·1%		2179

Table 2. Analysis of differences between short-listers in mean assessments and variances of assessments. Arrows between pairs of short-listers indicate that their means are different on a Scheffé test at the 0·05 level of significance, the arrow pointing from the dove to the hawk. Columns at the right-hand end indicate the overall significance of differences between means on a one-way analysis of variance, and the proportion of variance accounted for by those differences (eta squared). The column at the extreme right-hand end indicates the significance, using Cochrane's C test, of differences between variances, and of the ratio of maximum to average variance (range)

Scale	Response categories and expected proportions: Short-lister B C A D	Differences between means: % var	Significance	Differences between variances: Range	Significance
'O-levels'		0·4%	*	1·132	*
'A-level results'		3·0%	**	1·164	NS
'A-level predictions'		0·9%	**	1·046	NS
'A-levels'		0·7%	**	1·089	NS
'Interests'		10·2%	***	1·693	***
'Contribution to school'		2·7%	***	1·494	***
'Achievement'		2·1%	***	1·676	***
'Contribution to community'		3·9%	***	1·155	**
'Referee's report'		9·8%	***	1·360	***
'Potential'		2·8%	***	1·250	***
Short-lister's decision		0·7%	**	1·209	***
Dean's decision		0·2%	NS	1·062	NS
Final decision		0·1%	NS	1·071	NS

NS – not significant; $*P < 0{\cdot}05$; $**P < 0{\cdot}01$; $***P < 0{\cdot}01$

maximum variance over the average variance. Once again there are significant differences, but they are relatively small in that at most the maximum variance is 69% greater than the average variance.

Taken overall these results suggest that differences in means and variances have little actual effect upon the final decision reached. Although the Dean's decision and final decision are derived from the individual short-listers' judgements, and there are small differences in means and variances between these short-listers, there are no significant differences in rate of Dean's and final decisions when these decisions are analysed in relation to the particular short-lister who made the initial judgements. The use of a single person finally to assess all *pro formas* therefore compensated for the small differences between hawks and doves.

Analysis of correlations

The essential information for analysing the relationship between the separate items which are being assessed during short-listing is the Pearsonian correlation matrix of each of the items with every other item. Table 3 shows a typical such matrix for the 1986 judgements of short-lister A. The two variables assessing A-levels have been collapsed into a single composite variable (see above). Correlation matrices for short-listers B, C and D, and for short-lister A in 1981, are shown in the Appendix, in keeping with the principle that full matrices should be reported so that other researchers may investigate the consequences of structural assumptions other than the ones we have chosen to make in this paper (Kenny 1979).

Our previous exploratory analysis of the judgements of short-lister A had found three separate factors, which we identified as 'Academic ability', 'Interests' and 'Contribution to community', and these were, of necessity, orthogonal. For the present study we have used that analysis as the basis for a confirmatory factor analysis, in which a similar three factors were extracted, but allowed to be oblique (i.e. correlated).

A brief account of LISREL has been presented elsewhere (McManus & Richards 1986b), and several detailed accounts have also been published (e.g. Kenny 1979; Long 1983; Everitt 1984). Essentially the program allows causal relationships to be modelled between latent (or hypothetical) variables which are thought to underlie actual or measured variables. These relationships are expressed diagrammatically, latent variables being shown within circles, and measured variables within rectangles. Arrows between variables indicate direct causality (if straight and single-headed) or mere association (if curved and double-headed). Unattributable error variance is indicated by unattached arrows. Given a particular model then the path coefficients corresponding to each arrow can be estimated mathematically (and the following analyses used a maximum likelihood criterion). Path coefficients can be conceptualized as similar to beta coefficients in multiple regression, or factor loadings or correlations in conventional factor analysis. All coefficients are standardized since it is a correlation matrix which is being

Table 3. Correlations between the various items for short-lister A, during short-listing of 1986 applicants ($n = 623$; above diagonal). Below diagonal are shown the residuals from the model described in Fig. 1 and Table 4

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) 'O-levels'	—	0·561	0·333	0·396	0·418	0·127	0·497	0·524	0·446
(2) 'A-levels'	0·163	—	0·322	0·367	0·448	0·104	0·553	0·667	0·497
(3) 'Interests'	−0·002	−0·073	—	0·625	0·554	0·290	0·500	0·557	0·529
(4) 'Contribution to school'	−0·001	−0·101	−0·004	—	0·675	0·265	0·633	0·634	0·591
(5) 'Achievement'	0·029	−0·011	−0·009	0·008	—	0·239	0·597	0·626	0·542
(6) 'Contribution to community'	−0·052	−0·107	0·053	−0·015	−0·019	—	0·291	0·302	0·339
(7) 'Referee's report'	−0·004	−0·038	0·002	0·044	0·020	0·025	—	0·807	0·708
(8) 'Potential'	−0·022	0·023	0·015	−0·007	−0·003	0·013	−0·004	—	0·752
(9) Decision	−0·017	−0·049	0·030	−0·001	−0·027	0·000	0·020	0·003	—

analysed. The adequacy of a model is tested by examining residual differences between actual values in the correlation matrix and values predicted from the model. A conventional χ^2 statistic can be calculated, although this is invariably 'significant' with large samples, and an adjusted 'goodness of fit index' is generally recommended as a criterion of fit (Joreskog & Sorbom 1983). If a model is adequate then standard errors may also be calculated for each of the coefficients. If inadequate the model may be substantively altered by addition of paths, and re-estimated.

The path model was constructed to be as simple as possible, while still clearly derived from the basic principles obtained from the exploratory analysis of McManus & Richards (1984b). The three latent variables were allowed to be intercorrelated with one another (an oblique factor structure), and initially each measured variable was assumed to load only on a single latent variable, according to substantive considerations of the meaning of the measured and latent variables. Thus 'Academic ability' was assumed to be related to O- and A-level results, and 'Potential' and 'Referee's report' were felt also to be principally concerned with academic ability. In contrast it was felt that the latent variable 'Interests' should relate to the measured variables 'Interests' and 'Contribution to school'. And the latent variable 'Contribution to community' could necessarily only be related to the single measured variable 'Contribution to community'. The measured variable 'Achievement' was more problematic. Fitting of models in which it was only related to either 'Academic ability' or 'Interests' produced unsatisfactory fits. However, since the original definition of the variable explicitly said 'either or both special achievement in *any* activity and all-round achievement, including academic work' (McManus 1985, p. 158), the variable was therefore allowed to load on both 'Academic ability' and 'Interests'. As far as possible it was felt to be desirable to fit the same qualitative model to each of the five data sets, allowing differences between short-listers to appear only in quantitative estimates of parameters: that condition was readily satisfied.

Figure 1 shows the final path diagram relating the measured and latent variables for the present data. The middle column shows the three latent

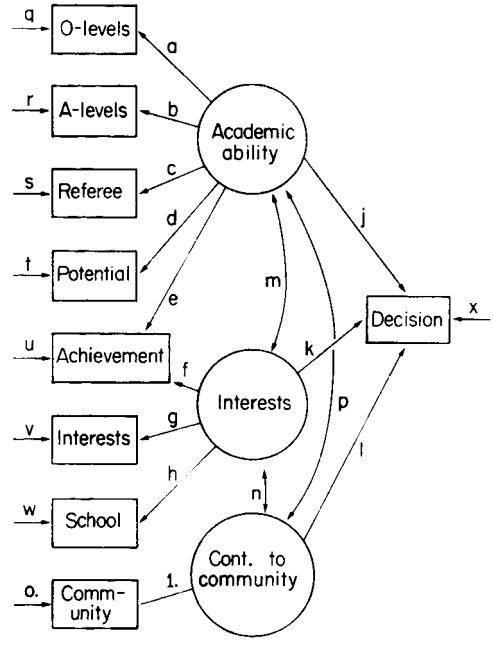

Figure 1. Structural diagram indicating the relationships between the nine measured variables (eight assessments and the final decision) and the three latent variables. Path coefficients are shown alongside paths as letters, and estimates of their values may be found in Table 4.

variables derived from the previous study. These latent variables represent the perceptions of the short-listers concerning three major attributes of the candidates, and on the basis of them a final decision is made, indicated by the rectangle on the right. The three latent variables cannot be measured directly, but manifest entirely through the set of eight judgements made about each candidate (indicated on the left-hand side).

The path model was estimated separately for the data from each of the four short-listers, and separately for the two sets of data from short-lister A. The coefficients estimated are indicated symbolically alongside paths in Fig. 1, and their estimates are shown in Table 4, in the order A (1981), A (1986), B, C and D. Adjusted goodness of fit indices for the five data sets were 0·909, 0·903, 0·863, 0·897 and 0·899 respectively, and root-mean-square residuals were 0·040, 0·040, 0·053, 0·050 and 0·040 respectively, indicating good fits between data and model. Table 3 and

the tables in the Appendix show residual values for each of the correlation matrices for the final models described here. The only consistent pattern visible in the matrix of residuals is an association between O- and A-levels, suggesting the possibility of a halo effect, candidates with good O-levels being felt to have better A-levels when these are assessed. Such a link could readily be accommodated into the path model by means of an association between the errors of the O- and A-level measures, and would probably result in a better fit of the model. Standard errors of loadings of manifest upon latent variables were typically of the order of 0·03, and those of correlations between latent variables of the order of 0·02, so that almost all coefficients shown in Fig. 1 are highly significantly different from zero.

Detailed examination of the path coefficients in Table 4 is of some interest. Comparison of the judgements of short-lister A during 1981 and 1986 suggest that the structure of his judgements has changed only very little during that period, the major change being that the assessment of 'Achievement' is more dependent upon 'Academic ability' in 1981 and upon 'Interests' in 1986. Similarly comparison of short-listers A, B, C and D in 1986 suggests that each is relating the three latent variables in a similar fashion to the eight measured variables. The most important difference between the short-listers is in the intercorrelations between the latent variables. In particular short-lister D has high correlations between the three latent variables (0·867, 0·711, and 0·742), suggesting that they are hardly differentiated, and that candidates tend therefore to be globally rated as simply good or bad. In contrast short-listers A, B and C are clearly delineating 'Interests' and 'Contribution to community', and 'Academic ability' and 'Contribution to community'. All short-listers except C show high correlations between 'Academic ability' and 'Interests', and the possibility must be raised that this reflects a genuine correlation

Table 4. Shows parameter estimates for the model of Fig. 1 for each of the four short-listers. All estimates are significantly different from zero at the 0·05 level except for those marked †

	Short-lister				
	A (1981)	A (1986)	B	C	D
Factor loadings					
a	0·626	0·581	0·575	0·508	0·607
b	0·365	0·685	0·649	0·584	0·602
c	0·829	0·863	0·735	0·871	0·837
d	0·924	0·940	0·963	0·947	0·977
e	0·500	0·156	−0·013†	0·181†	0·332
f	0·272	0·649	0·800	0·584	0·594
g	0·791	0·729	0·814	0·750	0·944
h	0·839	0·863	0·884	0·751	0·970
j	0·673	0·669	0·623	0·663	0·642
k	0·173	0·126	0·232	0·087†	0·216
l	0·018†	0·092	0·094	0·021†	−0·101
Correlations between latent variables					
m	0·728	0·791	0·624	0·883	0·895
n	0·356	0·324	0·165	0·422	0·729
p	0·364	0·308	0·236	0·516	0·737
Errors for measured variables					
q	0·608	0·662	0·669	0·742	0·632
r	0·867	0·530	0·579	0·658	0·637
s	0·312	0·256	0·460	0·241	0·299
t	0·089	0·117	0·072	0·102	0·046
u	0·478	0·394	0·372	0·440	0·184
v	0·375	0·468	0·337	0·438	0·110
w	0·297	0·256	0·219	0·437	0·060
x	0·336	0·349	0·334	0·433	0·409

within candidates rather than within short-listers' perceptions, so that it actually is the case that academically good applicants are also the ones with greater interests, although the data of the present study cannot separate actual differences between candidates from differences in the perceptions of short-listers. The decision of each short-lister is principally based on 'Academic ability', although all short-listers also use 'Interests' in making a decision, and to a lesser extent short-listers A, B and C also use 'Contribution to community'.

Discussion

Since short-listing, or pre-selection, is so important during medical student selection it is essential that the process should be adequately described, and should satisfy certain minimal statistical criteria. This paper, in conjunction with another on the reliability of the judgements made by short-listers (McManus & Richards 1989) provides such descriptions.

Short-listing is not peculiar to medical student selection, but occurs during personnel selection in general, to reduce the size of the 'mountain of applications' (Keenan 1983), with numbers typically being reduced by 50% or more (Herriot 1982, p. 58). Although the problem of graduate selection is in many ways akin to that of medical student selection, there is at least one important difference: Herriot (1982), in his comprehensive review of the processes of graduate selection, emphasizes that to a large extent employers are willing to accept that the intellectual ability and training of graduates are already satisfactory, whereas this assumption is one that cannot be made during selection of medical students, the majority of whom apply straight from school. Herriot points out that 'the application form is in principle a source of the most valid predictors of job success yet discovered: biodata' (p. 59), whereby 'biodata' is meant a vast range of 'historical and verifiable pieces of information about an individual', typically about background, training, qualifications, experience, etc. These items are reported very reliably, in the sense of being repeatable, although they are not always accurate, particularly if an applicant stands to gain from minor distortions, such as concerning previous salary. Wingrove *et al.* (1984) have examined the role of over 300 such biodata variables in predicting success during graduate selection in several industries, and found that typical correlates were with educational, work and leisure achievements, although there were variations between organizations and selectors in their use of this information. Keenan & Scott (1985) found that leisure activities were not an independent predictor of success in a less well-controlled study. Biodata predict success well over periods of a year or so, although they become less good predictors over longer time periods. Herriot also reviews the literature on 'reference reports' and concludes that both reliability and validity are typically low, validity being described by Muchinsky (1979) as 'from unacceptable to mediocre'. Once more such conclusions may not generalize well to the specific case of medical student selection where references on UCCA forms are provided typically by head teachers with much experience of the students and less motivation for giving inaccurate opinions. Herriot also reviews the literature on the problems which can arise in the use of rating scales during assessment, reporting Saal *et al.*'s (1980) conclusion that frequently there are 'halo effects', whereby there is a lack of clear differentiation between separate items, a high score on one carrying over onto another, and 'central tendency' and 'restriction of range', whereby assessors tend not to use the extremes of scales. It can be pointed out in partial mitigation that since on most measures of ability people are approximately normally distributed, so one would expect such a distribution to emerge during rating, with few candidates at the extremes: and this has been recognized by researchers such as Jones (1984) who recommended to short-listers that they should use five categories in the approximately normal proportions of 10, 20, 40, 20 and 10%. The present paper shows that short-listers are clearly capable of making recommendations in such proportions. Herriot also emphasizes the frequent research findings that there are many characteristics of raters which affect the ratings that they make; and Landy & Farr (1980) have raised the question whether indeed the factor structures found in assessments reflect 'behavioural patterns of ratees or cognitive constructs of raters'? Herriot & Wingrove (1984) have examined the

actual processes by which graduate recruiters have made pre-selection decisions, by transcribing their comments as they 'thought aloud' while reading an application form. Decisions were a weighted function of positive and negative evaluations and inferences, with greater weight being placed on items near the end of the form, suggesting a degree of information overload. Significant differences were found between selectors, although it is not clear whether these were of substantive as opposed to statistical significance.

Short-listers are generally capable of making judgements of UCCA applications in terms of a range of separate criteria (i.e. there is relatively little 'halo effect'), and are capable of spreading those judgements so that they broadly coincide with prespecified criteria (i.e. there is not much restriction of range). More importantly there are only marginal differences, of no practical importance, between short-listers in their tendency to place judgements towards one end or other of the scales. That is, differences between 'hawks' and 'doves' are minimal, and hence applicants can be assured of the fairness of a selection process in which each application is assessed in detail by only a single short-lister. These minimal differences are themselves removed by having a single person who broadly reviews all applications, to ensure similar proportions short-listed. Applicants can also be reassured that judgements made are reliable, both within and between short-listers (McManus & Richards 1989).

Our previous study had found that PR (short-lister A) in fact extracted three statistically independent categories of information from an UCCA form. This study has extended that analysis by fitting a confirmatory factor analytic model to those data, and confirming that the same model can be fitted both to PR's 1986 data, and also to the data from short-listers B, C and D. Such long-term stability and structural stability between short-listers suggests that indeed the extraction of at least three independent dimensions is a reliable process. Of some interest, however, is the finding that those three dimensions are strongly correlated for short-lister D (who was short-listing for the first time). The implication of this halo effect is that this short-lister was not able, for one reason or another, to make judgements with the same degree of discrimination as the other short-listers (one other of whom was also carrying out the process for the first time). The strong implication is that some short-listers may be better at the task than others, reflecting either greater care and effort or greater experience. The regular monitoring of short-listers' judgements, coupled with regular feedback, should improve the quality of the process.

Neither this study nor our other study (McManus & Richards 1989) attacks the problem of validation of selection procedures. This is not because the problem is unimportant, but because it is exceptionally difficult to approach. A valid selection procedure uses assessments at time T to predict behaviour at time $T + t$, where t may be several decades or more in the case of medical student selection. Furthermore, at time $T + t$ there must be satisfactory criteria for determining which individuals are indeed successful — the 'good doctors' that all selectors are looking for — but such criteria are difficult to define operationally and with general agreement. In the case of medicine the problem is made more difficult by the broad range of specialties which graduates may enter, so that selection at medical school entry cannot hope systematically to find the *specific* characteristics optimal for all of them. For the various specialties it is rather the case that subsequent further selection will take place for entry into each specialty, all of which have their own criteria for validity (and attempts to look at these are being made in at least one specialty, anaesthesia [Reeve 1984]). Nevertheless, despite all such difficulties, at least one strong conclusion is derivable from theory; if assessments do not show reliability then they certainly cannot have validity. Our present work shows not only that selection criteria for short-listing are reliable, but also that the structure of the judgements is similar in different short-listers. Although those findings cannot guarantee validity for our measures, they are at least compatible with it, and do not logically preclude it; to determine whether validity is indeed present requires long-term longitudinal studies, in which we are at present engaged.

Acknowledgements

We are grateful to Drs A. W. Boylston, B. P.

Curwain and G. H. Tait for their willing co-operation with this study, to Mrs R. Boyd and Ms C. Richards for their administrative help, and to the Economic and Social Research Council for financial support.

References

Anon. (1983) ***SPSS-X: User's Guide.*** McGraw-Hill, New York.

Everitt B.S. (1984) ***An Introduction to Latent Variable Models.*** Chapman and Hall, London.

Herriot P. (1982) ***Down from the Ivory Tower: Graduates and their Jobs.*** John Wiley, Chichester.

Herriot P. & Wingrove J. (1984) Decision processes in graduate pre-selection. *Journal of Occupational Psychology* **57,** 269–75.

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

Jones A. (1984) A study of pre-assessment centre candidate shortlisting. *Journal of Occupational Psychology* **57,** 67–76.

Joreskog K.G. & Sorbom D. (1983) *LISREL V and LISREL VI: analysis of linear structural relationships by maximum likelihood and least squares methods.* Department of Statistics, Uppsala.

Keenan A. (1983) Where application forms mislead. *Personnel Management* **15,** 40–3.

Keenan A. & Scott R.S. (1985) Employment success of graduates: relationships of biographical factors and job-seeking behaviours. *Journal of Occupationl Behaviour* **6,** 305–11.

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

Landy F.J. & Farr J.L. (1980) Performance rating. *Psychological Bulletin* **87,** 72–107.

Long J.S. (1983) *Confirmatory Factor Analysis.* Sage, Beverley Hills.

McManus I.C. (1985) *Medical students: origins, selection, attitudes and culture.* MD Thesis, University of London.

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. Shortlisting and interviews. *British Medical Journal* **289,** 1288–90.

McManus I.C. & Richards P. (1985) Admission to medical school. *British Medical Journal* **290,** 319–20.

McManus I.C. & Richards P. (1986a) Prospective survey of performance of medical students during pre-clinical years. *British Medical Journal* **293,** 124–7.

McManus I.C. & Richards P. (1986b) Admission for medicine in the United Kingdom: a structural model. *Medical Education* **20,** 181–6.

McManus I.C. & Richards P. (1989) Reliability of short-listing in medical student selection. *Medical Education* **23,** 147–51.

Muchinsky P.M. (1979) The use of reference reports in personnel selection: a review and evaluation. *Journal of Occupational Psychology* **52,** 287–97.

Reeve P. (1984) Selection of anaesthetists: is there a better method? In: *Quality of Care in Anaesthetic Practice* (ed. by J.N. Lunn), pp. 231–263. Royal Society of Medicine and MacMillan, London.

Saal F.E., Downey R.G. & Lahey M.A. (1980) Rating the ratings: assessing the psychometric quality of rating data. *Psychological Bulletin* **88,** 413–28.

Wingrove J., Glendinning R. & Herriot P. (1984) Graduate pre-selection: a research note. *Journal of Occupational Psychology* **57,** 169–71.