Does performance improve when candidates resit a postgraduate examination?

I. C. McMANUS

Academic Department of Psychiatry, St Mary's Hospital Medical School, London

Summary. Luck plays some role in passing any examination. When candidates pass a postgraduate examination at the second, third or subsequent attempt is it because their knowledge has truly improved or because they have at last been lucky? In this paper a simple model, requiring knowledge only of the pass rates at resits, and of the reliability of the examination, is applied to the MRCGP examination of the Royal College of General Practitioners. Candidates increase their true ability before second and third attempts at the examination, after which ability declines.

Key words: family practice/*educ; *educ, med, grad; *educational measurement; clinical competence; probability; models, statistical

Introduction

Passing examinations requires both skill and luck. A well-prepared and knowledgeable candidate may perform poorly because of minor illness or other distracters, and generally ill-informed candidates may just happen to be asked questions to which they have prepared answers. An examination with a good *discrimination* will show a strong relationship between a candidate's true ability and the probability of success, so that chance factors will be relatively unimportant.

Many candidates fail postgraduate examinations at their first attempt (Anon. 1990), and then some time later will resit the examination, perhaps repeating the process several times. Since chance processes influence the likelihood of passing an examination, it is therefore important to ask whether candidates who pass an examination on, say, the third attempt, are actually more *knowledgeable* than when taking the examination for the first time, or whether they have simply been *luckier*. Although both candidates and examiners are tempted to attribute eventual success to an increase in ability, that attribution is not necessarily correct. There are many board games in which one starts by throwing a six with a dice; however, when a six is eventually thrown, perhaps at the third or fourth attempt, one should not infer that eventual success is due to an improved ability at controlling the throw of the dice. Chance alone eventually results in the criterion being achieved.

This paper describes a straightforward model of success and failure in candidates repeatedly sitting a postgraduate examination, and fits the model to data available for the Royal College of General Practitioners' membership examination (MRCGP), a detailed description of which has been provided by Godlee (1991). The model could in principle be extended to a range of other questions, such as assessing the extent to which candidates who fail examinations are ill-prepared (Hardy 1990), perhaps by taking the examination too early.

Methods

The Annual Report to the Council of the Royal College of General Practitioners contains the figures shown in Table 1 in which are given the pass rates of candidates taking the MRCGP examination for the first, second, third, fourth and fifth or subsequent occasions, the data being accumulated across a total of 14654 candidates. The same report also shows that the pass rate for

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

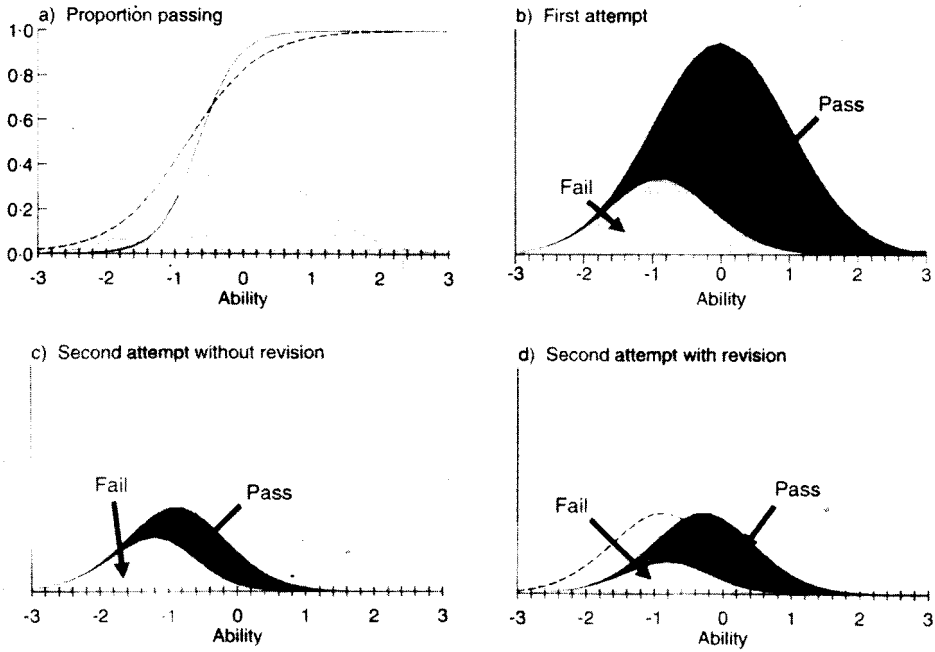

Figure 1. (a) The shaded area shows the distribution of ability scores of first-time candidates for an examination, with a mean of zero and a standard deviation of one. The dashed line shows the proportion of candidates passing the examination for parameters of $a = 1{\cdot}55$ and $b = 2{\cdot}00$ (model 1 in text), and the solid line shows the proportion of passes for parameters of $a = 2{\cdot}31$ and $b = 3{\cdot}53$ (models 2 and 3 in text). (b) The candidates taking the examination for the first time are divided into those passing (dark shading) and those failing (light shading), in relation to their ability. The probability of passing is calculated using the parameters of model 1. (c) The candidates who failed the examination on the first occasion take it on a second occasion, and again are divided into those who pass (dark shading) and those who fail (light shading). (d) The candidates who failed the examination on the first occasion (and who at that time have the distribution of ability shown by the dashed line) carry out additional study and therefore shift their ability distribution to the right, resulting in a higher proportion of passes (dark shading) and a lower proportion of fails (light shading) than in Fig. 1c.

It should be noted that at any particular ability level the proportion of candidates passing at any attempt is constant, and is the same as that shown by the dashed line in Fig. 1a.

the MRCGP examination remained almost constant from 1973 to 1987; although the model to be described can be applied to examinations for which the pass rate shows secular changes, it is more straightforward when the pass rate remains constant. Calculations have been carried out by numerical simulation using the SUPERCALC spreadsheat program on an IBM microcomputer.

Statistical model

The model described here is an adaptation of the logistic version of the Item Response Model (Goldstein & Wood 1989) for relating test performance to an underlying latent trait. It is assumed that the true knowledge (or latent ability) of the population of candidates taking an examination for the first time is normally distributed, and, without loss of generality, it is arbitrarily assumed to have a mean of zero and standard deviation of one (Fig. 1a). Examination performance (classified as pass or fail) is related to latent ability through a logistic regression, in which a candidate's probability of passing the examination, p_i, relates to the *difficulty* of the examination (a) and the *discrimination* of the examination (b). (Technical details of the statistical method are given in the Appendix.) Figure 1a shows the probability of passing the examination in relation to latent ability, for several values of b, the discrimination parameter. As b increases so

Table 1. Proportion of candidates passing the MRCGP examination at each attempt, and the predicted proportions from three different models of the data. 95% confidence intervals for the parameter estimates are shown in parentheses

	a Attempt	b *n*	c % pass	d Model 1	e Model 2	f δ_j	g Model 3
Parameter *a*				−1·55	−2·31		−2·31
				(−1·48; −1·61)	*		*
Parameter *b*				2·00	3·53		3·53
				(1·91; 2·11)	*		*
	1	12 647	72·0	72·1%	72·0%	–	71·9%
	2	1397	46·5	45·3%	30·4%	0·327	46·6%
						(0·275; 0·382)	
	3	325	35·0	31·9%	18·6%	0·160	34·8%
						(0·061; 0·254)	
	4	129	21·0	24·7%	13·5%	−0·050	21·0%
						(−0·199; 0·095)	
	5	156	9·0	20·3%	10·7%	(−0·427; −0·027)	

*No 95% confidence interval since parameter not estimated from data but determined externally.

the curve becomes steeper and there is a better discrimination of low ability candidates from high; good examinations are therefore associated with high values of *b*, being better able to discriminate truly good candidates from poorer candidates. As the difficulty parameter, *a*, increases, so a higher proportion of candidates at each ability level will fail the examination, and less will pass; 'difficult' examinations are therefore associated with higher *a* values. The ideal combination of *a* and *b* is one in which the steepest part of the functions shown in Fig. 1a occurs at the level of true knowledge deemed necessary for candidates to pass. Readers unfamiliar with logistic regression may find it helpful to note that if the curves of Fig. 1a are replotted against the logit of the proportion passing (see Appendix for description of the logistic function) then the curves become straight lines, equivalent to those normally found in conventional linear regression, and described in terms of their slope (discrimination) and intercept (difficulty).

Model 1. Resitting of an examination is straightforwardly modelled by considering all of those candidates who fail an examination on the first occasion and then allowing them to have a second attempt without any improvement in latent ability. Figure 1b shows for a discrimination parameter, *b*, of 2·00 (and $a = 1{\cdot}55$), the proportions of candidates of each latent ability who will pass or fail on their first attempt; and Fig. 1c shows the proportions of those who fail on the first attempt who will pass or fail on the second attempt (assuming that the examination were to be taken again immediately, without additional preparation). The process can be repeated for several resits and gives the data shown in column d of Table 1; *a* and *b* for model 1 are calculated to provide the best overall fit to the data (maximum likelihood estimates — see Appendix). This model, in which candidates do not improve at all in their true ability between resits, provides a reasonable fit to the data of Table 1, thereby demonstrating that eventual passing of an examination *can* be the result of an examination with a relatively poor discrimination, coupled with chance factors, and without any true improvement in candidates' ability between resits. Nevertheless, the data are not fit perfectly, since the value of the χ^2 goodness of fit test is 27·52 with 3 df ($P < 0{\cdot}001$).

Model 2. In model 1, the discrimination parameter, *b*, was estimated indirectly, from the data themselves, to provide a good fit to those data. However, *b* can also be estimated directly from a knowledge of an examination's reliability. The α coefficient of reliability (Cronbach 1951) for the MRCGP examination has been cited as 0·81 (Mulholland & McAleer 1990). This coefficient is effectively the correlation between scores on parallel versions of an examination (Ghiselli *et al.* 1981) and can be construed as the test–retest

correlation for the scores of the same candidates repeating the examination on two separate occasions. This is also equivalent to the tetrachoric correlation calculated from the proportions of candidates overall who would pass or fail the examination on both occasions. Values of b of 3·53 and a of −2·31, predict that 63·5% of candidates pass on both occasions, 19·5% fail on both occasions, and 17·0% pass on one occasion and fail on the other, thereby giving a tetrachoric correlation of 0·81 (equivalent to the α coefficient quoted earlier), and a pass rate at first attempt of 72%, as is known to be the case for this examination. These values of a and b are substantially outside the 95% confidence intervals for those estimated in model 1 (a: −1·47 to −1·61; b: 1·91 to 2·11). These improved estimates of parameters a and b, derived from the known reliability of the MRCGP examination, are shown as the solid line in Fig. 1a. Column e of Table 1 shows the expected pass rates for resit candidates for model 2 which uses these parameters. Now it is apparent that such a model alone does not fit the data (χ^2 goodness of fit test = 187·69, 4 df, $P << 0·001$), due to more candidates passing on second, third and fourth occasions than predicted by the model. In order to make the model fit it is necessary to take account of true changes in performance between repeated sittings of the examination.

Model 3. The models described thus far have taken no account of improvement in a candidate's latent ability between successive attempts at the examination; that is, candidates carrying out additional work between failing the examination and their subsequent attempt, and thereby increasing their true ability. Such additional study can be modelled by increasing the latent ability of each candidate sitting the examination for a second time by an amount δ_2, by an amount δ_3 for those sitting for the third time, and so on. In Fig. 1d the candidates who have failed the examination on the first attempt, and who then have ability scores shown by the dashed line, carry out additional study which increases their overall ability to that shown by the two solid distributions, resulting in a higher proportion of passes than that occurring in Fig. 1c. Using the values of a and b calculated for first time candidates in model 2, the values of δ_2, δ_3, δ_4 and δ_5 may be estimated so that they give the best fit to the data of table 1. Column f in Table 1 shows the estimated values of the δ coefficients for each resit, and column g shows the estimated pass rates for each group of candidates using those δ values. It can be seen that in order to fit the data it is necessary to postulate a relatively large improvement in the candidate's latent ability scores between the first and second attempts at the examination (0·327, about one-third of a standard deviation of ability), followed by a smaller additional improvement in ability before the third attempt (0·160), and then decreases in true ability before the fourth attempt (−0·050) and the fifth attempt (−0·215). A χ^2 goodness of fit test necessarily shows a perfect fit for this model since there are as many parameters as there are data points, but 95% confidence intervals for the estimates can still be calculated (see Table 1 and Appendix).

Discussion

A straightforward model is proposed of candidates taking postgraduate examinations on repeated occasions. It makes several assumptions: that the latent abilities of candidates are normally distributed; that candidates resitting examinations are a random subset of those who have failed; and that a linear logistic regression equation describes the relationship between latent ability and the likelihood of success in examinations. None of these is an exceptional or unrealistic assumption given the educational literature in general, and all are open to empirical testing given more data. If those assumptions are not satisfied then the model may readily be altered to produce a more sophisticated model.

Fitting of models 2 and 3 does require that an external estimate is available of the reliability of the examination, in this case an α of 0·81, and hence if that value were inaccurate then other parameter estimates would change. The dependence of the conclusions upon the size of α can be estimated by using a value of 0·7, rather than the actual value of 0·81; the estimates of δ_2 to δ_5 are then 0·185 (0·115; 0·246), 0·110 (−0·016; 0·226), −0·125 (0·057; −0·322) and −0·330 (−0·078; −0·619), 95% confidence intervals being shown in parentheses. The broad pattern remains the same as that in Table 1: candidates improve between the first and second attempt, and deteriorate

between the fourth and fifth attempts, although the effects are reduced in magnitude. The overall interpretation is therefore not unduly sensitive to the estimate of the reliability of the examination.

For the MRCGP examination it is clear that candidates *do* improve in true ability between repeated attempts, particularly between the first and second attempt at the examination. There is a lesser degree of improvement before the third attempt at the examination, and then, presumably because of examination fatigue and boredom, ability appears to decrease before the fourth and fifth resits. The results would seem to justify the practices of Colleges in setting an upper limit to the number of attempts which candidates may make at an examination (and without which it has been reported that candidates may sit up to 20 or more times (Hatch 1990)).

The model described here is capable of further development given more data. In particular if groups of candidates are compared (such as those sitting an examination very soon after qualifying, as compared with those sitting it later after qualifying) then it should be possible to assess whether candidates attempting the examination early are indeed ill-prepared, as has been suggested (Hardy 1990). A further development of the model, through multiple logistic regression, would allow the assessment of the effect of a range of background variables upon true ability, and upon improvement between resits.

Fitting of the model requires not only that pass rates are known for repeated attempts at an examination, but also that the α coefficient, or some other coefficient of reliability, is available for the examination. Alpha coefficients should therefore be reported routinely for all examinations, so that their reliability of their discrimination between weak and strong candidates can be assessed.

At present the model has only been fitted to data from the MRCGP examination. In an attempt to obtain further data a draft of the paper was sent to the Colleges administering the larger postgraduate examinations. Only a minority replied: the Royal College of Surgeons of England, although keen to help, did not keep its data in a readily accessible form (although a new computer system should allow such analysis); similarly the Royal College of Physicians and Surgeons of Glasgow was sympathetic but stated that manual data extraction would be difficult; and although interested in the model, the Royal College of Psychiatrists had recently changed its examination and therefore felt analysis to be premature. At a meeting of the Royal Colleges of Physicians administering the MRCP(UK) examination it was agreed that the data at present were incomplete, and not suitable for analysis, but that suitable data would be collected over the next 2 years. No reply was received from the Royal College of Pathologists, the Royal College of Obstetricians and Gynaecologists, the Royal College of Radiologists, or the College of Anaesthetists.

Royal College membership examinations are an important career step for many junior doctors, are expensive for those doctors to take (McManus 1991), and potentially are an important source of income for the Colleges. It is therefore necessary that the nature, function and validity of the examinations are available for public inspection, in order to reassure both the public as a whole, and examination candidates, that the examinations carried out by the colleges are properly carrying out their stated task. The present analysis would not have been possible without the data that were so readily provided by the Royal College of General Practitioners. It is hoped that in the future other colleges will also be able to make their data available; as Godlee (1991) has stated, 'the confidence with which the RCGP opens its doors to outside scrutiny sends an unmistakable challenge to other colleges to do the same.'

Acknowledgement

I am grateful to the Royal College of General Practitioners for providing the data on which this study is based.

References

Anon. (1990) Editorial: Examining the Royal Colleges' examiners. *Lancet* **335,** 443–5.

Cox D.R. (1970) *The Analysis of Binary Data.* Chapman and Hall, London.

Cronbach L.J. (1951) Co-efficient alpha and the internal structure of tests. *Psychometrika* **16,** 292–334.

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

Godlee F. (1991) MRCGP: examining the exam. *British Medical Journal* **303,** 235–8.
Goldstein H. & Wood R. (1989) Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology* **42,** 139–67.
Hardy K.J. (1990) Examining examiners [letter]. *Lancet* **335,** 731.
Hatch D.J. (1990) Examining the examiners [letter]. *Lancet* **335,** 916.
McManus I.C. (1991) Membership examinations and Royal College finances. *Lancet* **337,** 414–15.
Mulholland H. & McAleer S. (1990) Examining examiners [letter]. *Lancet* **335,** 731.

Appendix

The basic data to be fitted consist of the 5 × 2 table containing frequencies of candidates passing and failing the examination on the jth attempt, $j = 1,5$, where P_j and F_j represent the actual numbers passing and failing, and p_j and f_j ($p_j + f_j = 1$) represent the proportions predicted by the model. Maximum likelihood estimates of parameters (Cox 1970) were found by finding those values which maximized the log likelihood function, L:

$$L = \sum_{j=1,5} (P_j.\ln(p_j) + F_j.\ln(f_j))$$

95% confidence intervals of parameters were estimated as those values which decreased L by a value of 1 (i.e. a two-unit decrease in the support function).

The relationship between an individual's score on the latent trait of ability, x_i, and the probability of success in the examination, p_i, was modelled through the function:

$$\text{logit}\ (p_i) = b.\chi_i - a$$

where the logit (logistic) function takes the form:

$$\text{logit}\ (q) = \ln(q/(1-q))$$

ln() representing the natural logarithm (i.e. to base e). The latent trait was assumed *a priori* to have a mean of zero and variance of unity in the overall population. It should be noted that the difficulty parameter, a, is assigned a negative sign in the regression equation so that higher values indicate than an examination becomes more difficult.

Received 22 *March* 1991; *editorial comments to author* 11 *July* 1991; *accepted for publication* 6 *September* 1991