The Statistician (1998)
47, *Part* 2, *pp.* 291–321

Organizations selecting people: how the process could be made fairer by the appropriate use of lotteries

Conall Boyle†

University of Central England, Birmingham, UK

[*Read before* The Royal Statistical Society *on Wednesday, July 9th, 1997, the President*, Professor R. N. Curnow, *in the Chair*]

Summary. Organizations select people to receive benefits in a way which is efficient to them but may not be fair to those selected or rejected. This paper elaborates on the concept of fairness—that it should be efficient, not waste the efforts of the candidates; that it should treat as equals all those who are not measurably different; that the process of selection should avoid bias and corruption. Lotteries have been used in the past partly to avoid corruption. Some examples of lottery-type selection remain today, such as juries. This paper examines the case for the deliberate introduction of a lottery as part of the selection process to approximate to the uncertainty in measuring the merits of the candidates. The advantages of such a lottery, particularly where decisions are devolved down to the community level, are discussed.

Keywords: Efficiency; Fairness; Lottery; Organizational selection processes; Random selection

1. Introduction

1.1. Edgeworth's aleatory proposal

It was Edgeworth (1888), the eminent Victorian economist and statistician, who suggested that degree classification at Cambridge was rather a lottery. He went further and produced a neat formula to describe the probability that an individual candidate would be allocated to the wrong grade by being given, say, a first class instead of a merited second:

$$P_r = \frac{1}{2}\left\{1 - \theta\left(\frac{p}{0.477q}\right)\right\} \tag{1}$$

where θ is a normal probability, p is the probable error in the marks and q is the number of marks which the candidate is short of the dividing line between classifications.

Edgeworth had no difficulty accepting that examination marking was not determined precisely but was subject to measurable variability. He then went on to ask himself the question: why not decide the borderline cases by a lottery? He goes on to describe how it would be easy to contrive a solemn conclave of Fellows who would settle doubtful cases by drawing lots. But this stage, Edgeworth states, is superfluous:

> 'A public examination is already a sort of lottery of the graduated species, one which the chances are not equal, but are better for the more deserving …. It is a species of sortition infinitely preferable to the ancient method of casting lots for honours and offices.'

†*Address for correspondence*: School of Property and Construction, Faculty of the Built Environment, University of Central England, Perry Barr, Birmingham, B42 2SU, UK.
E-mail: conall.boyle@uce.ac.uk

 0039–0526/98/47291

Edgeworth (1890) returned to this subject 2 years later, perhaps prompted by a comment from a Mr Elliot on his earlier paper, pointing out the anomaly that with examinations for entry to the civil service 'serious differences of income and position could turn upon differences of marks which are largely or altogether accidental'. In his 1890 analysis of civil service entrance examinations, Edgeworth seems to have agreed with Elliot, that these examinations

> 'impose hardship on those just outside the gates of Paradise . . .. The general recognition of the element of chance in examinations would mitigate the disappointment.'

He then goes on to contradict his earlier view about selection by a graduated lottery. The candidates should be segregated by what he describes as a 'light' examination. They would then be given lottery tickets in proportion to the examination score. Commenting on this proposal, Edgeworth states that the state would not lose out since, *ex hypothesi*, in the long run the same proportion of really deserving candidates would be appointed. The benefits of such a selection were twofold: the sense of injustice felt by the candidates would be mitigated and the public would be alerted to what Edgeworth called 'the aleatory character of examinations' ('*aleatory*: depending on contingencies: used of the element of chance in poetic composition, etc. [L. *aleator*—a dicer, *alea*—a die]' (*Chambers English Dictionary*, 1990)).

1.2. The proposal: that deliberate randomization be introduced as part of selection

It is the hypothesis of this paper that selection of people by organizations could be made fairer by the deliberate introduction of an element of randomization, e.g. a lottery, as part of the process. The extent to which a lottery should be used in the selection decision should reflect the likely errors in the criteria used to select. (This will be made clearer through an example in Section 4.)

The proposal to use a lottery, or some other form of randomization, in serious decision-making is not new. The scientific approach to managerial and military decision-making has thrown up a whole range of situations where the logical decision is to toss a coin. What is new in the last two decades, according to Goodwin (1992), is the revival, in theory at least, of the notion of using some form of deliberate randomization to decide on the distribution of goods. This derives from the 'burgeoning of rational choice theory', and theoreticians in the fields of economics, politics and moral philosophy have advocated the use of a lottery for such a distribution.

This theory can lead to some extreme, not to say whimsical, suggestions. Goodwin described what she calls a mythical future Utopia called Aleatoria. This would be run using a total social lottery which decides, randomly, on all aspects of life. Other commentators like Broome (1984) and Elster (1992) agreed that a lottery could form a very limited part of the distribution process.

This paper attempts a practical addition to the theory of the use of lotteries, using the insights of statistics. Other theoreticians tend to deal in absolute concepts, treating the measurement of, say, 'merit' as something which in theory can be calculated exactly. They are aware of measurement errors but tend to look for situations where a lottery could be the sole arbiter. Statisticians have a more subtle approach, knowing that all observations should be reported with both a degree of certainty and uncertainty (due to error). The basic statistical–theoretical justification for the use of a lottery is that it recognizes and compensates for the errors in measurement in the selection process. A further contribution of statistics which is widely recognized is the understanding that a lottery is also a bulwark against bias.

1.3. A note on the use of the word lottery

I shall use the word lottery because it is widely understood, but any sound method of randomization could be substituted: this could include rolling a die, spinning a roulette wheel, or, biblically,

casting lots. It is usual in statistics text-books to introduce the concept of a *fair* die, meaning one which will show each of its faces with equal probability. In addition we would seek, as in the testing of pseudorandom number generators, to ensure that no predictable patterns or sequences occur. A randomization process which fails to satisfy these criteria is said to have *bias*. Note that the statistical–technical words fair and bias have other meanings in a social context, which will become apparent later. For convenience the description *lottery* will be applied from now on to denote any statistically fair process of achieving randomization.

1.4. Organizations which do the selecting

This paper is concerned with the non-market-based distribution of benefits to individuals, and how organizations deal with the process of selection of individuals for such benefits. The reasons for the existence of bureaucratic selection outside the usual market mechanisms include the following.

(a) Firstly, there is collective provision, where the market is held to have failed to distribute goods in an equitable manner. Over a wide range of goods, including housing, schooling and medical care, public, tax-funded provision is made for deserving individuals. This process requires public officialdom such as the civil service to follow rules laid down by government, sometimes allowing them to exercise their own discretion.

(b) Secondly, there is the establishment of firms within the economy, to capture the efficiency gains of working co-operatively. It was Galbraith (1987), p. 287, who pointed out that

> 'the reality ... is ... the dominant highly visible role in the modern economy of the great enterprise ... economics will deal increasingly with the internal and external dynamics of the large firm ... the future theory of the firm, if it is to have relevance, will of necessity be a theory primarily of bureaucratic structures and organisation'.

The administrative apparatus set up to manage firms like Imperial Chemical Industries or International Business Machines is strikingly similar to the civil service. Examples of selection by bureaucratic methods used by firms include the selection of suppliers and the choice of individuals for employment within the firm.

It is the way in which bureaucrats operate in both sectors and how this might be changed that forms the core of this paper. Clearly the bureaucracy is capable of performing its task of distribution; what is at issue is whether it could perform more effectively in promoting the welfare of the population or the firm that it is meant to serve.

2. Efficiency and fairness

Scientific experimenters often state that 'a sample was drawn from a pool (population) of data subjects'. No-one asks whether the members of the pool liked this procedure. So long as the sample yields good results for the experimenter, it is deemed to be useful. Selecting people is different. Selectors should be efficient, picking the best person for the job, the neediest family for a house or the cleverist pupil for a school place. But how does the process of selection look from the point of the candidate, the person on whom selection is exercised? Efficiency and fairness are the two sides of this decision. Selectors should act in accordance with the best interests of the organization, be it the government or the firm which employs them. An 'efficient' selection process could hence be described as a process which maximizes the benefit to society in the case of government or profit in the case of firms. This last is similar to the utility maximizing criterion used in economics.

'Fairness' could be described as the way that the process is viewed from the point of view of the individual who is being selected or rejected. A popular definition of fairness is that people should be treated 'on their merits' or 'according to their needs'. An operational definition of fairness will be described later in Section 2.2. Edgeworth, as noted earlier, was aware of the potential conflict of interest between the selectors and those selected—there is a need, in the public interest, to select the best people to be civil servants, while not being unfair to all those who constitute the pool of applicants. Efficiency for the organization should have primacy, but, if different methods of achieving equivalent efficiency are available, choose the one which is fairer all round.

2.1. Efficiency in selection

Efficiency (or utility as it is sometimes called) has been discussed in general statistical terms by Aitchison (1970). The allocation of resources by central government to health or education authorities has also been the subject of many papers of the Royal Statistical Society's journal Series A (see Derbyshire (1987), Sheldon and Carr-Hill (1992), Straf (1981), Copas (1993) and Carpenter (1983)). In these papers, the consideration is what might be termed the general good, which reduces to the simple criterion of achieving the maximum output with the minimum cost. By so maximizing, government and firms can decide rationally on the effectiveness of supplying extra resources, or of switching resources.

A good example of this strategy is given by Carpenter (1983) in his paper on the reduction of post-natal infant mortality. The technique used was to calculate risk factors for each infant, and to calculate a score based on this. Extra care was then targeted on those babies scoring above the 'care line', excluding those with a lower risk score. As a result the overall infant mortality rate was reduced. Sensible calculations can flow from such exercises about the likely benefit of increased spending.

What is left out of these calculations is a consideration of *individual* need. A baby scoring just below the care line may have been wrongly allocated: is it fair that she be deprived of this care? How reliable is the conversion of points scored into mortality risk? There will always be some margin of error, so that some babies and their parents will have the trauma of unnecessary medical intervention, whereas others might have missed out on a life-saving intervention. From the administrator's perspective these two factors will balance out, and the general good will be maximized according to the best knowledge that is available at the time. The very healthy and very sickly babies will be correctly selected and treated. Many close to the borderline will be treated *unfairly.*

2.2. Fairness

To a statistician, fairness has resonances associated with random sampling. A fair die is one which can be expected to show any one of its six faces after a roll with a probability of $\frac{1}{6}$. There are different methods of drawing a representative sample from a population, but all depend on some form of randomization. Simple random sampling operates so that each member of the population has an equal chance of being selected. Samples which have not followed strict randomization may exhibit bias. Fair, bias and equal chance are ideas that will be familiar to all statisticians, as well as the implications of such techniques—that there will be occasions when a single sample produces a freak result, but in the long run such variation will even out. This is a subtle and complex process which is not well understood outside the profession.

Whereas 'fair' may have a strict scientific meaning in statistics and sampling, fairness is much less clear cut in the social sciences. A useful definition which will be adopted here is that

> 'Fairness means that relevantly like cases should be treated alike ... it could be argued that even where there are relevant differences, people should be treated alike'

(Elster (1989), p. 113). In the real world, fairness can only be an aspiration, but, following the ideas of Deming (Neave, 1990) for a strategy of incremental improvement towards this goal, we can hope to produce selection procedures which are *fairer*.

Elster elaborates on requirements of fairness in respect of the processes of selection and suggests how they might be judged.

(a) The selection process should be efficient: this would encompass the efficiency considerations above, as seen from the organization's viewpoint. Individuals should not be required to waste effort needlessly: open competitive examinations may be quick and cheap for the selectors but may waste effort, for example, in the learning of dead languages. It is especially costly for those who fail at such examinations. Filling in lengthy application forms and having to travel extensively are also costly to individuals; selection procedures should aim to reduce these search costs and perverse incentives for the candidates.

(b) Another requirement is what statisticians would call 'not significantly different'. It is convenient for adminstrators to have clean cut-off points in scoring systems or examinations. But this will be unfair because actual scores are subject to random variation, and scores are somewhat uncertain indicators of need, merit or future performance. Seen from the organization's viewpoint this does not matter; efficiency is maintained, and on average the better or the more needy are selected. The individual may not see it that way; he may have been turned down for a job, house or school place, believing that less qualified people have been accepted. The size of this border zone group, of those scoring close to the cut-off line, will vary according to the accuracy and reliability of the selection methods. How this variability might be measured and embodied in the selection procedure will be explained more closely in the case-study presented later in Section 4.

(c) The selection method should avoid bias: it used to be legal to discriminate against people on grounds of gender or race. This is no longer so, and such barriers to entry are banned. Manifestations of partiality are similarly outlawed: it would be illegal in the UK to advertise 'No Irish need apply'. The equal opportunities policy of my university affirms that

> 'No student or member of staff receives less favourable treatment on the grounds of *gender*, *race*, *sexual orientation*, *age* and *disability*',

which is a good example of the statements issued by organizations, and recognizes that avoiding partiality is more than the removal of active discrimination. Despite this, selection committees can be extremely biased (Morgan *et al*., 1982).

The unintentional prejudices of individual selectors are also significant. Efforts are made to ensure that selectors eschew bias on grounds of gender, race, sexual orientation, age or disability (the five grounds mentioned above). But the list of grounds for bias could be extended: investigations have uncovered a wide variety of personal attributes which may disadvantage individuals, despite having equal merits or needs—

'*heightism*', i.e. tall people are more successful than short people are (Economist, 1995a);

'*lookism*', i.e. selectors in interviews are biased towards prettier candidates;

'*hairism*', i.e. bald people are disadvantaged relative to hairy competitors (Guardian, 1995a);

'*weightism*', i.e. fat people are seen as less worthy than their slender counterparts.

This list could be extended, almost without limit. Readers may object that these are trivial forms of discrimination compared with racism or sexism. Investigations have shown otherwise, that real hardship is encountered by those who are perceived as less worthy.

(d) The selection process should be free from corruption: 'It is the natural condition of every bureaucracy to be corrupt' stated Lord Bancroft, former Head of the UK civil service (Guardian, 1995b). The bribing of public officials to give preference to some applicants over others is not wholly absent in the UK, which rightly prides itself on the probity of its officials. Other countries, often in the Third World, are not so fortunate. In its extreme form in Renaissance Italy, it seemed that no-one could trust anyone else outside the family to act honestly.

2.3. How organizations select

Selection methods could be classified as the following.

(a) Total discretion: the selector chooses without having to explain why. Selection panels which exercise their judgment in this way may be highly subjective.
(b) Strict rules: for example promotion is by seniority. This could be interpreted as a surrogate measure of merit, namely experience, together with a chance element of date of appointment.
(c) A points scoring system: where measures of need or merit are calculated and added up and the highest score wins. Although this appears fair and objective, deciding on what items to score and what weight to give to each is not simple. Rigging the points system to favour some groups is commonplace.
(d) A test or open competitive examination: this is normally related to the purpose for which the selection is being made, sometimes perverse, like requiring knowledge of antique Greek for selection as a colonial governor.

Actual selection systems are usually a combination of some of these methods. Thus a test may be followed by an interview (delegated patronage). Scoring systems may include an element of bureaucratic discretion to allow 'queue-jumping' for what are deemed to be especially needy cases. Whether these methods are efficient *and* fair depends on many factors. Which criteria are used and how they are established are important. The training and personality of the selectors and the constraints imposed on his or her freedom of action ('discretion') are probably the most important factors. We sometimes expect a superhuman degree of impartiality and prescience from our selectors, which is at variance with reasonable expectations. Rather than expect the selectors to acquire impossibly high standards, we should seek fair and efficient selection processes which do not require subjective judgments about fellow human beings.

2.4. External pressure to be fair

The onus is on the organizations themselves to be efficient in their selection procedures. They may also feel that ensuring they act fairly towards applicants is a goal that is worth pursuing, in the public interest. Firms may find it beneficial to act fairly towards potential and actual employees to encourage staff morale and hence productivity. External pressure on organizations can be exerted through legislation. Laws to promote fairness and equality started with the elimination of barriers to entry, such as the rules which prevented women from attending university. In more recent times it was felt that women and ethnic minorities were being excluded from jobs, housing and education through the bias of the selectors. Discrimination of this sort was outlawed by various Acts of Parliament and in the UK backed up by permanent commissions such as the Equal

Opportunities Commission and the Commission for Racial Equality. Manifestations of discrimination are outlawed, and organizations are encouraged to practise equal opportunities.

Laws which punish transgressions of equal opportunity legislation have proved of some use in alleviating unfairness. Selectors in firms and government agencies at least try to appear non-discriminatory. In reality, although there has been some progress, fairness towards all has not been achieved through legislation. Efficiency may also have been compromised, with candidates being chosen, not because they are deemed best, but because they help to fill an informal quota. Time spent by firms in ensuring compliance or dealing with alleged cases of discrimination may also reduce efficiency.

The use of *statistics* has been important in the discovery and attempted elimination of unfairness. Monitoring of the numbers of students by age, race and sex is commonplace in most colleges; employers are urged to monitor their employees in the same way. Measurement of the *outcomes* of selection processes is widespread.

Affirmative action and *positive discrimination* seek to direct benefits towards disadvantaged groups (and by implication away from other 'advantaged' groups).

Overall the removal of barriers to entry has been beneficial to groups which had previously been less well housed or educated, or had lower status jobs or incomes. The later attempts via monitoring and affirmative action have not produced anything like the same improvements for disadvantaged groups (Economist, 1995b). It could be argued that trying to fix the outcomes of the selection processes would always be an uphill struggle. What is required is to ensure that the *process* of selection is genuinely fair, so that the outcome comes closer to reflecting the true worth of each candidate rather than any group label that they might carry.

3. Lotteries for decision-making

3.1. Use of lotteries in historic times

There are many references detailing the use of lotteries in the classical era (Headlam, 1933; Staveley, 1972). The Romans and especially the Greeks used the lot to decide a range of things, such as who should serve as leaders. For the Greeks a lottery was felt to be the essence of democracy. Sometimes the techniques employed for randomization were bizarre, such as the examination of the entrails of an animal. There was also an element of pagan religion, with the process said to involve the discovery of the will of the gods. This was the case for certain monotheistic faiths, e.g. Judaism and Christianity. The Christian bible makes several positive references to decisions which were made 'by casting lots'. When the Apostles needed to find a replacement for Judas they chose between Justus and Matthias thus:

> 'Thou Lord who knowest the hearts of all men, show of these two the one thou hast chosen . . . and they gave lots for them; and the lot fell on Matthias' (*Acts*, chapter 1, verse 23).

Renaissance Italy also made use of lotteries for the selection of their rulers at all levels. This tended to be associated with the more democratic phases of their history. Once the autocratic Medicis took over Florence, selection by lot was abandoned (Najemy, 1982). In Venice the ruling oligarchy continued to select among themselves through a complex system of juries and lots. Some elements of selection by lot persist: to this day the tiny republic of San Marino draws out the name of its two leaders (*capitani regenti*) from an urn containing 12 possible names.

It might be thought that the use of lotteries in the past is an example of, at best, irrationality or, at worst, benighted superstition. But lotteries served some useful functions: they helped to resolve conflicts between warring factions; they were proof against corruption and bribery; they were simple to operate, requiring little record keeping; the act of drawing out the name or names of the

chosen was often turned into an elaborate ceremonial which served a useful social bonding function. Jacob Burckhardt referred to this as *Der Staat als Kunstwerk*, implying that the schemes of governance they developed showed genius comparable with da Vinci, Michelangelo or Dante (Burke, 1972).

3.2. Present-day uses of lotteries in selection

The onset of the Enlightenment swept away nearly all use of lotteries in selection processes, to be replaced by universal suffrage and bureaucratic rule, with the bureaucrats controlled by the elected representatives. There still remain some vestigial elements of random selection.

(a) Juries in legal proceedings are drawn from the population at large. There is some disagreement whether the jury should be randomly selected or should be representative, including a proportion of various groups. The opportunity of challenging jurors allows some departure from pure random selection. This is a matter which is not fully resolved, and the belief is that it is wrong to delve too deeply into the traditional workings of the law (Guardian, 1989).
(b) Occasionally some minor aspects of administration and legislation will be decided 'by lot'. A dead-heat in an election is one example. A place in the queue for private members' Bills in the UK Parliament is decided by a ballot (lottery).
(c) Selection for the US 'green card' entry permit: 40 000 places (16 000 reserved for Irish people, north and south) were allocated by means of lottery in 1992 (Independent, 1992).
(d) Selection for 100 of the 173 places at a popular ex-grammar school in Burnley, Lancashire, is decided by a lottery, a system which has been in use for 13 years.

> 'Application forms are shuffled and numbered by one official, while another chooses numbers from a set of random number tables drawn up by computer and reads them out'
>
> (Independent, 1994).

There are surprisingly many examples of the use of lotteries, such as selection for military service in the USA, entry to medical school in the Netherlands or the allocation of social housing in Israel (Elster, 1992), but with the possible exception of juries these are not part of mainstream administrative procedures. A lottery seems to have been chosen where other methods would simply not have worked.

4. A case-study: selection at 11 years of age for school placement

This section explains how a lottery might have been used as part of a selection process, in a way which would retain the claimed benefits while being more fair to the applicants.

4.1. School selection by using an intelligence test

Discriminating between people on the basis of an intelligence quotient (IQ) test has had a troubled history. Relevations that Sir Cyril Burt, one of the pioneers of testing, may have faked much of his data on monozygotic twin studies casts a shadow over such endeavours. More controversially is the questionable use of statistical results which may lend credence to racism, in for example the best selling book *The Bell Curve* (Herrenstein and Murray, 1994). It might seem foolhardy to venture into such controversial territory, but I do so for a good reason: the testing of IQ to predict short run academic performance has been intensively studied, and a wealth of

accessible and reliable information is available. I know of no other test which has received so much scrutiny. Whether any alternative to IQ testing or indeed once-only selection at the age 11 years would be better or fairer, I leave to the judgment of the reader.

4.2. The old British eleven-plus test for grammar school placements

This section draws heavily on Vernon (1957), who reviewed the evidence for the effectiveness of the eleven-plus IQ test. A more recent publication by Gipps and Murphy (1994) covers some of the same ground but does not challenge any of the earlier figures which were produced concerning the accuracy and reliability of this test.

The objective of the British eleven-plus test was to measure the IQ of all children in the 11-year-old cohort within each local education authority (LEA). This could involve tens of thousands of school-children in a single authority (borough), so there was plenty of scope to establish fair and efficient procedures. On the basis of their scores on the test, a percentage of the pupils from the cohort, which ranged from 14% in Nottinghamshire to 60% in Merionethshire, were allocated to grammar schools, in the belief that they could benefit from an academic style of education.

The measure of success for the eleven-plus test was very simple: how well did the test predict the performance of the cohort 5 years later at the General Certificate of Education (national, public) examinations? The short answer is very well indeed, especially compared with alternative methods of selection and prediction. The alternative selection methods which were used included standardized tests in mathematics and English, ranking by teachers and special entrance examinations set by individual schools. A global figure for the reliability of IQ tests in predicting later examination scores was estimated by Vernon and his fellow workers at a correlation coefficient of 0.70. All other methods showed lower correlations.

The implementation of the eleven-plus test varied from one LEA to another. It was appreciated that the test was not perfect, and that a sharp cut-off point would result in the unfairness of candidates being wrongly allocated. For this reason most LEAs adopted a 'border zone' procedure, calling for further reports on candidates who fell just below the cut-off point. As time went on this border zone shrank, mainly for practical reasons. What was needed, according to one shrewd local councillor, was a test which was 'technically sound, administratively feasible and politically defensible' (Vernon (1957), p. 30). The IQ test seemed to be sound. For administrative and political reasons the border zone was progressively shrunk.

Vernon and his co-workers felt that this was wrong, that the border zone should be expanded. For the typical LEA which allocated grammar school places to about 25% of the 11-year-old cohort they suggested the following (Vernon (1957), p. 169): on the basis of the eleven-plus IQ test the top 5% should be automatically admitted, the lowest 50% eliminated. The remaining 45% should be subjected to further assessment by subject-based examination, teacher assessment and individual interview.

4.3. How an efficient and fair selection system might have been operated

In Section 3.2 it was proposed that a test which was fair to the applicant should satisfy four criteria one of which was that candidates should be treated as equal to the extent to which the test cannot reliably distinguish between them. It is this criterion of fairness that I shall use to try to establish the extent to which a lottery might have determined pass or fail on the eleven-plus test.

All the stages of the selection process involve error, that is an inevitable part of any measurement and prediction process. To be fair to the candidates the size of this error should be estimated using the best knowledge available. This should then be used to treat candidates 'according to their merits' or at least in proportion to their probable merit. This could be described, using the

language of statistical process control, as the operating characteristic of the test: if an LEA had decided that an IQ score of 115 was the cut-off, what is the probability that a candidate with a score of 100 *might* have achieved the cut-off value?

Errors arise at different stages of any test and create uncertainty over its predictive ability.

(a) *Marking errors* arise from slips and blunders made by the markers. These should not be great, especially if well-designed and electronically read response forms are used. An estimate of the likely error from this source can be found in Carpenter (1983) who calculated that about 3% of entries in a scoring scheme were erroneous. Since this can work both for and against the candidate, and some blunders will cancel out, a figure of 1% wrongly allocated will be taken.

(b) *Repeatability errors* relate to the candidate's performance, which can vary from day to day, and may be affected by varying physical or psychological factors. Although a small repertoire of well-validated tests was used they could not be identical in their performance, which again produced a problem of repeatability. Vernon (1957), p. 85, estimates that the variability from all these sources could result in about 10% of candidates being wrongly allocated.

(c) For *prediction errors* Vernon suggested that the IQ test, together with other supplementary selection procedures, could reliably predict academic peformance 5 years later with a correlation coefficient of up to 0.90. From this he deduced that about 15% would be wrongly allocated.

These three categories of error or 'wrongly allocated' can be combined. Since 'error' is a concept that is similar to variance, we could add these figures to produce an overall estimate of the 'power' (in a statistical sense) of the IQ test:

marking error	+	repeatability error	+	prediction error
1%	+	10%	+	15%

giving a total of 26% wrongly allocated. Some challengeable assumptions are involved in this

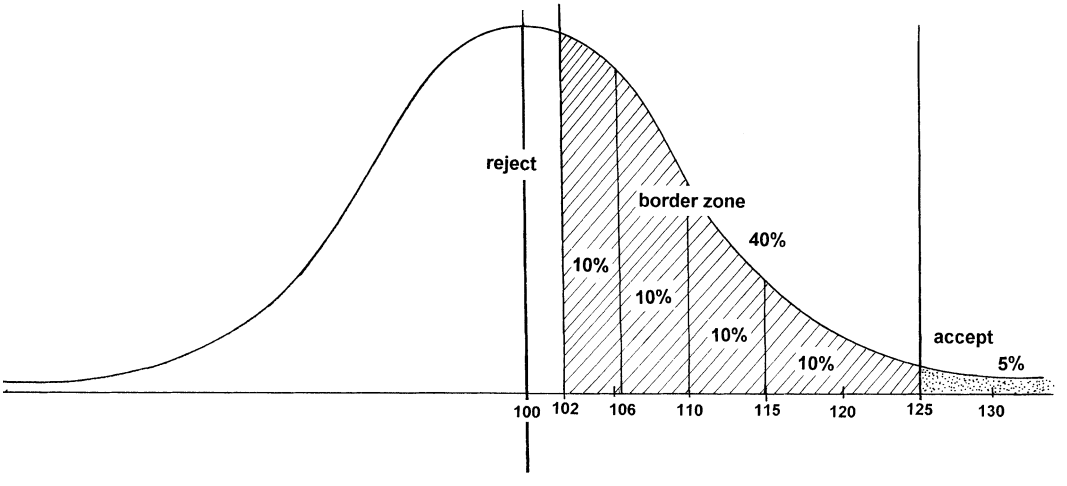

Fig. 1. Hypothetical *N*(100, 15) distribution of IQ test scores: the top 5% are accepted and the bottom 55% are rejected; the remaining 40% fall in the border zone and are segregated into four subgroups, each containing 10%, as shown

calculation, but it is encouraging that the result produces a border zone similar to that which Vernon recommended. Given that the typical proportion of grammar school places was for 25% of the 11-year-old group, we can attempt to translate this into values on the normal distribution.

Applying the 26% percentile border zone to a selection process which accepts 25% and rejects 75% produces the effect shown in Fig. 1. It is a matter of decision or further investigation whether to have equal percentages of candidates on either side of the 110 cut-off score or to take the lowest 26% of those above 110 and the highest 26% of those below. Clearly one system will advantage the high scorers; the other will reach down to a lower level of scores (cut off at 105 and 102 respectively).

Following Vernon the procedure to be followed in this case could be framed thus:

(a) the top 5% on the test score—automatic entry;
(b) the bottom 55% on the test score—eliminated;
(c) the intermediate 40% on the test score—enter the border zone lottery.

Within the border zone candidates should be selected in such a way that their chance of being selected rises with the score they have achieved on the test—the 'graduated lottery' described by Edgeworth previously. Fig. 2 shows a possible method, with the border zone divided into four subgroups. A little arithmetic and the use of a table of percentages under the normal distribution produce the cascade illustrated.

Candidates could be told their actual score; the draw could be a public event. This would have a salutary effect in communicating the aleatory (chance-dependent) nature of the process. Other advantages of such a lottery include

(a) inefficient 'cramming' for the test, which helps the well-off who can pay for it, would be less certain of conferring an unfair advantage, and so would be less used, and

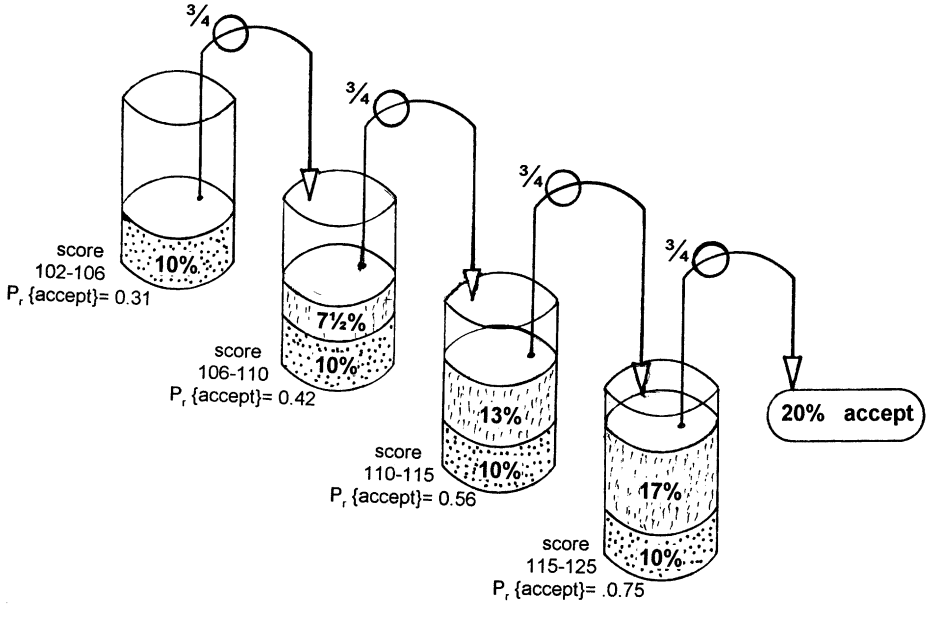

Fig. 2. Graduated lottery: candidates scoring between 102 and 125 form the border zone of the IQ test (see Fig. 1); they are split into four groups as shown; at each stage three-quarters of the candidates are selected randomly to go through to the next stage

(b) cultural bias within any test would not be eliminated, but good candidates thus disadvantaged would not be systematically excluded. Some would survive the lottery, even if less than on their inherent merits. This is not perfect but gives much improved chances over the strict cut-off system.

5. Conclusions and comment on lotteries for selecting people

5.1. The technical basis for using a lottery as part of selection

The case made in this paper is that the judicious introduction of a lottery into at least part of every selection process would enhance fairness and at the very least would not harm the efficiency of the selecting organizations. The extent to which a lottery should be invoked depends on the amount of uncertainty in the selection process and arises from the second criterion of fairness described in Section 2.2. To be fair to the applicant we should 'give them the benefit of the doubt'—look at the objective measurable criteria and ask 'What is the lowest possible score that *might* have been sufficient to achieve a job or an entry or a desired grade?'.

The candidates are divided into three groups:

(a) those who score well above the cut-off line are accepted without further process;
(b) those who score well below are rejected;
(c) those who fall 'close' to the cut-off line go forward to a lottery.

To calculate the size of the border zone depends on the accuracy of the scoring method. This needs to be evaluated. Whether a graduated or simple lottery should be used is a matter of taste and statistical sophistication.

It is important both for public confidence and to ensure reliable results that the draw be made in public and the results published. Lotteries can go wrong for quite innocent reasons as Fienberg (1973) pointed out: in the 1969 US selective service draft lottery, the mere 'stirring of capsules in a goldfish bowl' did not produce random ordering. More draftees were selected from January than from December because of the order in which they were put in the goldfish bowl. If the results of the draw had not been published this flaw might never have come to light.

5.2. The philosophical case for the use of lotteries as part of selection

The argument developed in Section 2 took fairness to the candidate and efficiency for the organization as the criteria which needed to be fulfilled. This section briefly reviews these and asks how selection which includes an element of random chance measures up.

5.2.1. Efficiency for the individual

A lottery is quick and cheap. Since the lottery-based selection process would generally be simple, this should mean less effort by the candidates in form filling, being subjected to extensive testing or maybe even a large amount of travelling. Some selection processes are based on need; to succeed, the applicant must show how distressed they are. This creates perverse incentives to exaggerate. A lottery would reduce this incentive.

5.2.2. Equality

Taking the description of fairness from Elster (1989) quoted in Section 2.2, that 'relevantly like cases should be treated alike', the technique described in Section 4.3 is no more than a statistican's

way of measuring 'not significantly different'. The main novelty is to take this from the viewpoint of the candidate, not the selector; hence fairness to the candidate.

5.2.3. Bias

No selection process could ever be devised that is totally free of bias and that treats everyone on their precise objective merits. What can be claimed for selection with a lottery is that it gives everyone a measurable chance, despite any bias, including for those for which it might be difficult to gain public sympathy. This represents an enormous advance on current attempts to alleviate a very limited repertoire of biases.

5.2.4. Corruption

Self-evidently if the dice decide then no-one can predict the outcome. Bureaucrats without any taint of corruption would be ideal, but when, as is proposed, decisions should be passed down to a vast range of local communities, lottery-based selection becomes an essential safeguard. Etzioni (1993), the arch-advocate of 'community' in all its forms, is aware of the dangers of local corruption and bias, although he offers no specific advice about how to protect against them. A lottery offers such protection.

5.2.5. Efficiency as seen from the point of view of the organization

The benefits of cheap and quick methods of selection are extolled by both Elster and Broome. It could be administratively much less irksome to select via a lottery. The mounting requirements for monitoring and recasting of selection procedures to comply with edicts to select in a non-sexist, non-racist etc. manner could be curtailed. The main anxiety of the organization would be that they were destined to receive an intake which was consistently worse on average. Edgeworth, as noted in Section 1.1, thought not, that *ex hypothesi* the same proportion of really deserving candidates would be appointed. This would often be because unmeasured criteria could be vital.

5.3. A caveat: the need for repeated lotteries

One feature of random processes is well known to statisticians: a single lottery is quite capable of producing a rogue result. For Elster (1992) this is enough to condemn the whole idea: if a lottery was used to select organ transplantees, a lottery could roguishly select a patient with a poor prognosis. This would be 'unacceptable' according to Elster; he suggests instead simple rules like 'always take the youngest'.

Although it may be true that a single lottery result may be freakish, in the long run a fair (in the statistical sense) lottery can be expected to yield nearly equal numbers of outcomes. For this reason, I would not advocate any process where all is decided on a single throw. Returning to the organ transplant case, a multiple-stage lottery could be contrived, with more tickets for those with a better prognosis and quotas for different subgroups. My main objection to simple rules is that they systematically discriminate against certain groups, depriving them of any hope of release. A lottery, however badly stacked against you, at least gives some hope of a successful outcome.

Another method of ensuring that candidates are not subjected to a single lottery is to disaggregate the outcome of each decision. The benefit being allocated should be the smallest practicable: a 1-year job contract for example. In addition to breaking down the goods to be allocated into smaller parcels, encouragement should be given to repeat and multiple applications. Thus each medical school should allow minimally qualified candidates to re-enter their

admission lottery once or twice. In this way the number of repeated lottery events for each person over a single lifetime could rapidly mount up.

5.4. Psychological barriers to acceptance of lotteries as part of selection

However rational and logical the proposal to run part of a selection process by means of a lottery may be, there will still remain deep-seated subjective revulsion towards the idea. This would come, I suggest, from people who *should* know better, and from those who *do not* know any better. By this I refer to the 'experts' on the one hand and the general public on the other.

The technocrats advise the bureaucrats, who prefer clear-cut solutions. It is difficult for a technical expert to explain about the uncertainty of his or her knowledge and much easier to claim that, where knowledge runs out, judgment or experience is enough to engender trust. The decision to use a lottery requires the technical expert not only to state and defend the extent of his or her knowledge but also the extent of non-knowledge about the outcome of a selection process. Are the experts ready to admit to and explain their limitations?

The general public at least has the excuse of being misled about the true nature of a lottery. Journalists frequently refer to some haphazard or uncertain process as 'a mere lottery'. The results of legal proceedings are often described thus. A more profound psychological anxiety has been identified by Elster (1989): the need to have an explanation, to seek a cause for every decision. Elster described cases where both natural parents fight for custody of their child. The evidence is usually highly subjective and unreliable, and the decision very finely balanced. Would this not be an ideal application for the toss of a coin? Elster rejected this, on the grounds that people prefer a pesudoreason for a decision, rather than be told 'it's a toss-up'. He even suggested that having an identifiable decision taker diverts the anger that the losing parent might feel for the winning parent onto the judge in the case. Elster may be correct in his judgment in this instance; that practical social work is best carried out by acting out a charade. I can only return to Edgeworth's conclusion and point out that using a coin toss to decide difficult cases would alert the public to the aleatory nature of selection decisions.

Acknowledgements

I would like to thank Paul Allin for encouraging me to pursue this offbeat idea and the referees for their helpful comments.

References

Aitchison, J. (1970) Statistical problems of treatment allocation (with discussion). *J. R. Statist. Soc*. A, **133**, 206–238.
Broome, J. (1984) Selecting people randomly. *Ethics*, Oct., 35–55.
Burke, P. (1972) *Culture and Society in Renaissance Italy 1420–1540*. London: Batsford.
Carpenter, R. G. (1983) Scoring to provide risk-related primary health care: evaluation and up-dating during use (with discussion). *J. R. Statist. Soc*. A, **146**, 1–32.
Copas, J. B. (1993) A formula for the allocation of resources based on uncertain predictions of need. *J. R. Statist. Soc*. A, **156**, 107–113.
Derbyshire, M. E. (1987) Statistical rationale for grant-related expenditure assessment (GREA) concerning personal social services (with discussion). *J. R. Statist. Soc*. A, **150**, 309–333.
Economist (1995a) Heightism. *The Economist*, Dec. 23rd, 21–26.
———(1995b) A question of colour. *The Economist*, Apr. 15th, 13.
Edgeworth, F. Y. (1888) The statistics of examinations. *J. R. Statist. Soc*., **51**, 599–635.
———(1890) The element of chance in competitive examinations. *J. R. Statist. Soc*., **53**, 644–663.
Elster, J. (1989) *Solomonic Judgments: Studies in the Limits of Rationality*. Cambridge: Cambridge University Press.
———(1992) *Local Justice: how Institutions Allocate Scarce Goods and Necessary Burdens*. Cambridge: Cambridge University Press.

Etzioni, A. (1993) *The Spirit of Community: Rights, Responsibilities and the Communitarian Agenda*. London: Fontana.
Fienberg, S. E. (1973) Randomization for the selective service draft lotteries. In *Statistics by Example: Weighing Chances* (ed. F. Mosteller). Reading: Addison-Wesley.
Galbraith, J. K. (1987) *A History of Economics: the Past as the Present*. London: Hamilton.
Gipps, C. and Murphy, P. (1994) *A Fair Test?: Assessment Achievement and Equity.* Buckingham: Open University Press.
Goodwin, B. (1992) *Justice by Lottery.* Hemel Hempstead: Harvester Wheatsheaf.
Guardian (1989) Random selection the essence of jury procedure. *The Guardian*, Aug. 18th.
———(1995a) Pates put bald men in a state. *The Guardian*, Aug. 29th.
———(1995b) An ethos up for sale by Lord Bancroft. *The Guardian*, Dec. 20th.
Headlam, J. W. (1933) *Election by Lot at Athens*. Cambridge: Cambridge University Press.
Herrenstein, R. J. and Murray, C. (1994) *The Bell Curve: Intelligence and Class Structure in American Life*. London: Free Press.
Independent (1992) Advisers aim to win with the US visa lottery. *The Independent*, Aug. 4th, 8.
———(1994) Parents challenge lotteries for school places. *The Independent*, May 31st, 4.
Morgan, C., Hall, V. and Mackay, H. (1982) *The Selection of Secondary School Head Teachers*. Milton Keynes: Open University Press.
Najemy, J. (1982) *Corporatism and Consensus in Florentine Electoral Politics 1280–1400*. Chapel Hill: University of North Carolina Press.
Neave, H. (1990) *The Deming Dimension*. Knoxville: SPC.
Sheldon, T. A. and Carr-Hill, R. (1992) Resource allocation by regression in the National Health Service: a critique of the Resource Allocation Working Party's review. *J. R. Statist. Soc*. A, **155**, 403–420.
Staveley, E. S. (1972) *Greek and Roman Voting and Elections*. London: Thames and Hudson.
Straf, M. L. (1981) Revenue allocation by regression: National Health Service appropriations for teaching hospitals. *J. R. Statist. Soc*. A, **144**, 80–84.
Vernon, P. E. (ed.) (1957) *Secondary School Selection—a British Psychological Society Inquiry.* London: Methuen.

Discussion on the paper by Boyle

Anne Hawkins (*University of Nottingham*)

Boyle addresses some complex issues in an extremely accessible way, persuasively arguing the case for including a lottery stage within selection procedures. I trust that readers will not be too diverted by the prospect of 'Mystic Meg' presiding over the recruitment of all future fast track entrants to the civil service. Several very serious points are made about shortcomings of more conventional selection procedures. If the ingredients are so unreliable, I wonder whether Boyle's proposed (later stage) lottery can salvage the process, or will we still end up with 'garbage in–garbage out'?

The Matching Education, Assessment and Employment Needs in Statistics project team has looked at graduate recruitment to posts involving statistical duties. Selection strategies are far from standardized. This is not the same as Boyle's point concerning the inherent variability of assessments that might be used within selection procedures. Job advertisements also vary greatly in how much detail they give. If we cannot, or do not, specify our selection criteria, we not only miss telling the world what statistics is (and hence what statisticians have to offer) but we also risk falling foul of Boyle's first two criteria—*efficiency* and *fairness*. It saves us from having to admit our fallibility as selectors, however.

Boyle proposes that subjectivity should be removed from selection processes. It is somewhat ironic, therefore, that he justifies the use of lotteries largely in terms of two criteria that he sees as being *subjectively* based. He also asserts that lotteries can be *seen* to have these qualities. However, when it comes to *subjective perceptions* about lotteries, seeing is not believing. There is a widespread lack of understanding about uncertainty (and its terminology). Belief in the likelihood of winning the lottery is subject to many misconceptions. If it was not, why would anyone enter the lottery? The expected outcome is not favourable ... but then how many people expect the expected outcome? There are also many subtle nuances in the perception of 'fairness'. Boyle's suggestion that one aspect may be concerned with wanting to feel that you have *some* chance, however minute, is interesting.

In a research context, I like to think of efficiency as meaning *more (or the equivalent) for the equivalent (or less)*—'more what' and 'less what', being negotiable. However, by implication, efficient research would be *valid* and *reliable*, i.e. it would be designed and conducted in such a way as to ensure a reasonable chance of producing the right answer, that could be replicated. I would see efficiency and fairness as having *objective external* reference points, as well as an element of forward planning. To Boyle, though, they are *subjective* perceptions on the part of the selectors and the candidates respectively, with more of a retrospective feel.

Boyle talks of the eleven-plus examination as a 'good' test because it has been much investigated, although I am not convinced that all these investigations were themselves that good. We concur though that the test was not wholly 'culture free'. Nor was it resistant to coaching. I would therefore question whether the eleven-plus test was 'technically sound' or 'politically defensible'. Boyle's suggestion that adding a lottery procedure counters such sources of *systematic* bias is appealing, but it becomes less attractive if those entering the lottery only do so because of their performance on a fallible test (in terms of its potential for bias). Vernon and colleagues were right to be concerned about borderline judgments, but the more fundamental question is whether the intelligence quotient really is a good measure of scholastic aptitude.

I endorse the author's criticism of assessment as being an imprecise art. I would, however, particularly draw attention to the inappropriateness of assessment instruments applied to statistical skills and understanding. Inevitably, when we use theoretical examinations to assess applied statistical skills, courses become more theoretical. Then we introduce project work to assess a candidate's applied skills, although we have done nothing to teach them such skills.

I am less sanguine than Boyle that advantages of lotteries can 'clearly' be seen unless we first give people a better understanding of the statistical processes involved. Boyle himself acknowledges that selectors may have difficulties in explaining uncertainty. He feels that a more widespread use of lotteries could raise the public's understanding of chance and uncertainty. There are few signs that the National Lottery has had this effect. I therefore remain unconvinced that educating the general public in 'things statistical' can be left to the chance mechanisms of incidental learning. The responsibility must rest with us.

Nevertheless, I want to say how very much I enjoyed this thought-provoking paper, and I congratulate Boyle on challenging perceived wisdom and the mechanisms at work in selection. It is therefore my pleasure, on behalf of us all, to engage in the 'elaborate ceremonial which serve[s] a useful social bonding function' (to quote from Boyle) and to propose a formal vote of thanks.

Barbara Goodwin (*University of East Anglia, Norwich*)

Conall Boyle's paper illustrates an interesting connection between philosophical principles and statistical analysis. For moral philosophers, there are certain archetypal cases where random allocation is the most ethical solution—e.g. situations of indeterminacy or where social justice is at stake. Elster and Williams have advocated distributive lotteries in situations of *indeterminacy*, where rational arguments cannot produce determinate outcomes. Such situations fall into the following three categories.

(a) People are so equal that we cannot choose between them. Elster call this *equi-optimality.*
(b) People are so different that they cannot even be compared. Elster calls this *incommensurability* (Elster (1989), pages 107–108).
(c) Situations where the decision is so momentous that human reason cannot make a choice, or human beings do not wish to take the responsibility for making a mortal choice. For example, one might toss a coin to decide which of two people to rescue from a fire where only one can be saved or (using Boyle's example) draw straws to see who should be thrown out of a leaking lifeboat. In Williams's words, there are situations where the use of a lottery is 'a reminder that some situations lie beyond justification' (Williams (1981), p. 18).

In all three cases, random choice would be the fairest solution from an objective standpoint; it would also save time and be more efficient (avoiding protracted efforts to make fine distinctions between individuals). Those chosen or not chosen would feel that they had been dealt with fairly and the selectors would be saved from making impossible decisions and exonerated from feelings of guilt. The question of indeterminacy relates to social justice, since situations where no rational choice can be made require a just decision procedure.

Random allocation can be used to promote social justice where people have approximately equal claims to scarce goods, or where their needs, capacities or other relevant attributes are roughly equal. I have argued elsewhere that a series of reiterated lotteries could produce a more just and egalitarian society under certain conditions (Goodwin, 1992). Boyle intimated a similar view in his presentation when he stated that 'Everyone should be subjected to lots of lotteries in all sorts of ways'.

Boyle's paper deals with both indeterminacy and justice. He advocates a more just distribution of goods or positions, e.g. in choosing pupils for selective schools or appointing candidates to jobs. He explores indeterminate situations, where candidates or applicants in a 'bordlerline zone' are

approximately equal (in statistical terms 'not significantly different') or where there is a measurable risk that they have been wrongly classified. In Elster's terms, Boyle is concerned with what we might term *approximate equi-optimality* of examination or job candidates in the borderline zone rather than with the incommensurability of the candidates.

Boyle advances two different claims simultaneously. In the eleven-plus example he argues statistically that there is a risk of 25% of candidates being erroneously classified. A lottery would not remove the error but would counteract it. Whether counteracting chance with chance works well is a question for statisticians rather than philosophers. Second, Boyle contends on the grounds of justice that the 45% 'borderline zone' candidates should not be denied all chance of entering a grammar school: they are not so markedly less talented that they should be automatically rejected. Boyle's two claims lead to the same conclusion—to random allocation for borderline candidates. I believe that the logic of his claims entails a lottery for *all* the candidates, for the following reasons.

(a) The statistical argument: any individual in the borderline zone may have been wrongly marked. But so (to a lesser extent) might any individual in the 'reject' zone or the 'accept' zone. (The examiner might transcribe 39% instead of 93% or vice versa.) Repeatability and prediction error could also affect the two extreme groups as well as the borderline group. A more general lottery would also randomize the self-confirming effect of eleven-plus selection. Should we not therefore conclude that all candidates should be entered into a weighted or graduated lottery such as Boyle illustrates in Fig. 2?
(b) The justice argument: as a matter of justice, perhaps all candidates should be given some chance in a weighted lottery so that no-one is debarred from success *ab initio* because of lack of talent. A weighted lottery, with higher chances for children with high intelligence quotient and lower chances for children with low intelligence quotient would prevent demotivation and disillusion for the latter. As Boyle says 'a lottery, however badly stacked against you, at least gives some hope of a successful outcome'.
(c) Against pride and despondency: Boyle also emphasizes that it is salutary for people to be aware of the chance-dependent nature of selection processes. But should not this salutary effect operate for both talented and untalented candidates as much as for borderline candidates? This also would suggest a wider lottery.

To conclude, there are sound arguments for using weighted lotteries in certain circumstances and they have sometimes been so used. Affirmative action could usefully be promoted by giving ethnic minority candidates extra tickets in any lottery to ensure that the number of such candidates awarded university places (for example) was proportionate to their percentage in the population. Boyle's paper gives an elegant demonstration of how such a weighted lottery could be devised in practical terms.

The vote of thanks was passed by acclamation.

Toby Lewis (*University of East Anglia*)

Mr Boyle's interesting paper is about making a process 'fairer'. What is *fairness*? Some think that it is obvious and needs no explanation; Mr Boyle rightly emphasizes that it is a complex concept. His view of fairness is, however, open to question. He uses (Section 4.3) a criterion where

> 'candidates should be treated as equal to the extent to which the test cannot reliably distinguish between them',

and states (Section 2.2) that the requirement

> 'is what statisticians would call "not significantly different" '.

Does he believe that there is no difference between

(a) no evidence at all for rejecting a null hypothesis (e.g. my mark is 59; yours is 59) and
(b) evidence for rejecting the null hypothesis existing but not reaching statistical significance (e.g. my mark is 59; yours is 61)?

In a recent book on assessment, 15 authors have written on different topics. I looked to see what they said about fairness, and I found some useful references. Cresswell (1996) asks

> 'What constitute meritocratically fair selections?'.

He discusses the relationship of fairness to *comparability of standards*, with an illuminating analysis of eight different definitions of comparable standards. Vincent (1996), writing on assessment in the workplace, says:

> 'In order to demonstrate fairness in the tests [which employers use for selection], ... thorough job analyses [need to] be conducted to identify that the constructs measured by the test are reflected in the abilities needed for success in the job'.

Similarly, Greaney and Kellaghan (1996), writing on the integrity of public examinations in developing countries, say:

> '[We need to be sure that] the marks or grades awarded candidates are directly related to the ability that is being measured rather than to irrelevant factors or uncontrolled conditions'

such as decision-making by lottery? As Elster (1992) says

> 'Although the randomness of lotteries may be superior to the potential capriciousness of discretionary selection, both are often perceived as more unfair than a system in which cases are judged publicly on the basis of easily verifiable criteria'.

From his different viewpoint Mr Boyle (Section 5.4), after magisterially rebuking many of us—

> 'people who should know better, and ... those who do not know any better'—

speaks of

> 'the aleatory nature of selection decisions'.

Assessments may indeed be *imprecise*; but does he really think that the resulting selection decisions are *aleatory*, in other words just matters of chance?

David J. Bartholomew (*Stoke Ash*)

This is one of the most thought-provoking papers which I have read and yet it leaves me wondering whether the kind of lottery proposed actually achieves its stated object. For example, if someone with a score in an admission test lower than mine were to be chosen by lot while I was rejected, would I think that I had been fairly treated?

The reason for my unease can be clarified by a simple modelling exercise in the spirit of Edgeworth. Suppose that suitability for treatment, higher education or whatever, is denoted by y and that it is scaled in such a manner as to render its distribution in the relevant population standard normal. Suppose, further, that the object is to select a proportion α of the population with the highest values of y, i.e. all those with $y > y_\alpha$ where $\Phi(y_\alpha) = 1 - \alpha$. We observe X on each individual which is an uncertain indicator of y. Since some individuals may be advantaged or disadvantaged in various ways, let X be a biased indicator and assume that for any candidate $X \sim N(y + b, \sigma^2)$ where b is the bias. This bias will doubtless vary in the population so let $b \sim N(0, \tau^2)$. (These particular assumptions are for illustration only: the point being made is quite general.)

It seems to me that the quantity which is relevant for selection purposes is

$$\Pr\{y > y_\alpha | X\} = p(X), \text{ say.}$$

For the model specified above

$$y|X \sim N\left(\frac{X}{1+v}, \frac{v}{1+v}\right)$$

where $v = \sigma^2 + \tau^2$. It easily follows that

$$p(X) = 1 - \Phi\left[y_\alpha \sqrt{\left(\frac{1+v}{v}\right)} - \frac{X}{\sqrt{\{v(1+v)\}}}\right].$$

The important point to notice is that $p(X)$ is a monotonic function of X. In other words, the larger my X the more likely it is that I belong to the target group. An element of lottery is already implicit in the random variation of X. The fairness of selection based only on the ranking of X resides in the monotonicity of $p(X)$. We could, of course, select individuals randomly with probability $p(X)$ but why should this appear fairer to the candidates?

The probability $p(X)$ *could* be used for a two-stage selection process using a rule which selects if $p(X) > 0.95$ and rejects if $p(X) < 0.05$ say. To calculate this probability we need to know σ^2 and τ^2 but, as the author points out, we may have some idea of the size of σ^2, at least, and rough approximations are all that are needed. A decision based on further evidence is surely fairer than adding further randomness to that which is already there.

Gerald Goodhardt (*London*)
I would like to give one example where a lottery method is used, and one where I think that it could be introduced with advantage.

In the Jewish religion, the Sabbath morning service provides the synagogue with the opportunity to honour some members of the congregation by giving them particular roles in the service. Some roles are more honorific than others, and whereas certain circumstances bestow priority, such as a bridegroom just before his wedding or the anniversary of the death of a parent, the choice of honorands is usually in the hands of the lay officials of the synagogue. A few years ago, I visited the synagogue in Madrid. On entering, each congregant was handed a numbered disc by the beadle. When the time came to distribute the honours, an official drew similarly numbered balls from a bag. This seemed to me a very fair procedure which removed any suspicion of favouritism. It rewarded congregants in proportion to the diligence of their attendance, without the need to keep records. This is important, since writing is forbidden on the Sabbath.

My second example concerns the allocation of grants by bodies such as research councils. As I understand it grant applications are first assessed by the system of peer review. However, this results in far more applications being judged worthy of support than there is money to fund them. The Economic and Social Research Council has reported that, in 1995, 522 applications were given the highest grade but only 225 could be funded. Therefore a second round of sifting is undertaken by, I believe, the various boards and committees. The criteria used in this round are not generally known, but it is widely believed that a major influence is the past record of applicants in successfully completing past projects. This is a reasonable risk reducing strategy, but it tends to lead to the creation of a limited circle of researchers, and to discourage new applicants from applying. My suggestion is that the majority of the money, say 80%, should be allocated as at present, but the remainder should be allocated by means of a lottery among the remaining top-graded projects. No-one can be aggrieved by this system since it cannot be known who would have benefited in the absence of the lottery. The winners in the lottery need not be identified either, unless it was felt desirable to publicize the method.

Mike Derbyshire (*Lancashire County Council, Preston*)
I can see advantages, from the viewpoint of organizations, in obtaining a more varied spread of successful applicants than might arise from traditional selection methods. However, as I understand the paper, one of the main thrusts of the argument is that the approach proposed would provide a fairer methodology from the point of view of the candidates; it would mitigate the sense of injustice felt by disadvantaged groups. To provide a practical context for discussion, in Table 1 I give (Lancashire County Council, 1997) data which describe the performance in the 1995 General Certificate of Secondary Education examinations of Lancashire children from various racial groups, in terms of the percentages obtaining five or more grades A* to C.

Asian boys and girls appear to be disadvantaged compared with white pupils. The effect of adopting the approach proposed by Boyle would be that some of the Asian children who failed to achieve the current threshold would be judged to have passed and that some who managed (against formidable

Table 1. Lancashire children obtaining five or more grades A* to C, 1995

Group	*Total number of children*	*% boys*	*% girls*
White	12496	35.5	43.3
Pakistani	807	22.5	18.7
Indian	376	29.1	37.6

Table 2. Lancashire children obtaining five or more grades A* to C, 1996

Group	*Total number of children*	*% boys*	*% girls*
White	10859	40.1	50.4
Pakistani	773	39.3	50.3
Indian	380	43.3	57.1

obstacles) to pass the present threshold would be judged to have failed. It is unclear to me that the gains for the first group would outweigh the loss, and sense of injustice, of the second group.

The situation is, however, more complex than these data indicate. In the following year, 1996, there was a general 'improvement' in performance, but the performance of Asian children surpassed that of white pupils, producing the data in Table 2 of percentages obtaining five or more grades A* to C.

We now have the bizarre position that, if we put our doubts aside and assume that Boyle's proposals would assist underperforming groups, then white boys would have been major beneficiaries in 1996. As more data become available, it will be interesting to see whether the various groups who suffer disadvantage have indeed learned how to win by the rules—at the same time as we are pondering how to change these rules fundamentally.

Roy Carr-Hill (*University of York*)

I shall comment—very briefly—on the ideological correlates and political implications of alternative methods of selection; the relevance or not of Boyle's proposal for resource allocation and the implications for 'rationing' in the National Health Service (NHS).

Selecting people is time consuming for selectors and very painful for those selected in an individualistic society which elevates financial success above worth. Any procedure which highlights the arbitrariness of 'success'—and, in addition, is less onerous for selectors—is therefore to be welcomed; Edgeworth (1890) was right in emphasizing the arbitrariness of selection procedures (compare Michael Young's *The Meritocratic Society*).

However, Boyle's contrast between a lottery and the current—nearly scientific, nearly statistical—methods of selection is incorrect; instead we should look to the Confucian method of designing job *desiderata* so that there would be only a very small number—and preferably only one—reasonable candidate.

In arguing for his proposal Boyle cites the problem of resource allocation. But the focus of *those* papers was how to weight groups in populations—usually, although not necessarily, geographically defined—to capture variations in need. Although superficially similar—in that the statistical manipulations which are the basis of those allocations are indeed subject to error (Martin *et al.*, 1997)—the focus is on the budget that is available for *all* health authority activities rather than with the selection or not of an individual for treatment.

But the suggestion has direct relevance to this problem of rationing in the NHS. Some economists argue that this should be on the basis of a preassessed capacity to benefit (expected gain in quality-adjusted life-years) relative to costs for the procedure and purportedly 'valid' measures of health quality are increasingly promoted (Ware *et al.*, 1993; EuroQol Group, 1990). There has been statistical criticism (Carr-Hill, 1989; Cox *et al.*, 1992) of the procedure in general and others have demonstrated specific biases and errors of measurement (Hunt and McKenna, 1991, 1993; Carr-Hill and Morris, 1991), and the substantial variability in the ratings which are used (Carr-Hill, 1991). The measurement of health quality is an inexact art: indeed, given the standard errors (Carr-Hill, 1989) most indexes proposed add hardly any information. The proposal for a graduated lottery (see West (1987)) will help to demonstrate the fallibility of these procedures, and to make the debate about rationing more democratic.

R. A. Kempton (*Biomathematics & Statistics Scotland, Edinburgh*)

The author considers efficiency and fairness in human selection procedures. An efficient selection procedure of fixed cost is defined as one which results in the appointment of candidates who will

provide the greatest expected benefit for the organization. A fair procedure is one which is perceived as treating all candidates according to their true merits.

A comparison with selection in other situations is of interest. In the development of crop varieties in the UK, Kempton and Talbot (1988) identified two distinct phases of selection. Firstly, individual plant breeders aim to select one or two potential new varieties from a large pool of genetically distinct plants. All such candidate varieties are then tested against current commercial varieties in statutory trials and, after further testing, the best performers are included on a 'recommended list'. For the first phase of selection, efficiency, measured by average genetic gain, is paramount and the concept of fairness does not exist (Finney, 1958). However, when comparing candidate varieties from several breeders, both efficiency and fairness need to be considered.

For an individual plant breeding programme which has fixed resources, there is a balance to be struck between the number of candidates assessed and the intensiveness of assessment. In practice, a multistage selection system is adopted, starting with a light screening of a large number of candidates followed by the selection of successively smaller numbers for more intensive testing. Taking this pattern to its extreme, Finney (1958) showed that, in some circumstances, the efficiency of selection could be increased by randomly selecting only a proportion of candidates for entry to the first stage. This contrasts with the author's proposal for random selection at the end of the testing process which will almost inevitably (*pace* Edgeworth) result in some loss of efficiency.

For statutory variety trials, as for selecting people, it is important that all participants perceive testing as fair, i.e. free of bias and with an acceptable level of random error. One measure of fairness used is the acceptance probability, defined as the probability that a candidate of known performance relative to the commercial standard will be accepted. An important question, which might have been discussed in greater detail in this paper, is the extent to which the need to demonstrate fairness compromises efficiency of selection.

Kenneth Evans (*Time and Tide Limited, Surbiton*)

Conall Boyle's suggestion of using lotteries in examinations close to the grade boundaries, on the grounds of greater fairness, is I think mistaken. If one candidate scores more than another, even by one mark, then it is at least slightly more probable that the first is the better candidate and it would be perverse to reorder the candidates in the grading.

Where the question is who should be recruited for a job, I think that his arguments are stronger but still mistaken, because the recruiter must be accountable to the organization. Still, the recruiter often does not know the right criteria for selection and often recruitment does not take into account the skills needed by the assembled team. In this context, Conall Boyle's arguments about administrative simplicity and freedom from corruption make interesting reading.

However, I do think that our speaker is onto something. A good analogy is with genetic algorithms. Here the problem is maximizing a function of many variables. The method is to start with a population of trial solutions and to use some rule for generating another population. The method is iterative and one keeps generating new populations until one is satisfied with the best solution obtained so far.

The distinctive step is to take two solutions from the previous generation, a 'father' and a 'mother', and to obtain a new solution by selecting the values of some variables, randomly chosen, from the father and the others from the mother. (Practitioners vary in how they implement these ideas.) There is also a mutation step in which a new solution is generated by random perturbations from a previous solution. The resulting algorithm works well. In my experiments, I have found the mix in Table 3 to be very effective.

Table 3. Rule of thumb for genetic algorithms

Source	*Approximate % of next generation*
Mating	70–80
An improvement formula	15–20
Random mutation	5–10
Total	100

The point for the present discussion is that the algorithm degrades badly, if there are no random mutations, even if we include a plausible hill climbing formula.

There is a lesson here—if we want to develop a learning organization, we need some random input. A proportion of nominations to representative bodies could be decided by lot. And, used in a small way, the idea could be the salvation of that most eccentric and quintessentially British institution: the House of Lords.

Paul Creighton (*Recruitment and Assessment Services Ltd, London*)

Against the background of criticism of recruitment selection that is evident in the paper (as presented) and the following discussion, I feel that I ought to raise my head above the parapet.

I am a Chartered Occupational Psychologist who designs and monitors the assessment system for recruitment to the Civil Service Fast Stream Development Programme. We are acutely aware of the issues of fairness, reliability and validity in assessment. Indeed, the British Psychological Society awards a professional qualification which is required for individuals to purchase, administer and interpret psychological tests. We are continually working to reduce both error and systematic bias. In our stated aim of 'fair and open competition with selection based on merit', we believe that the best approach is to ensure that we are assessing qualities which are important for good job performance and to design assessment methods which are as fair, reliable and valid as possible.

Rather than applying a lottery, our selection systems are arranged so that increasingly resource-intensive, but increasingly accurate, assessment methods are applied at successive stages. Many more candidates are passed to subsequent stages than there are vacancies, so those below the likely ultimate success boundary, at any stage, are given the chance to succeed under the more rigorous assessment at subsequent stages until the final stage is reached, when selection decisions are made with the fullest knowledge practically possible.

Nevertheless, the paper raises some important issues in recruitment selection. Major concerns in recruitment are the questions of 'cloning' and of how organizations can change and develop. It is possible to have a selection system which appears perfectly valid, i.e. a high correlation between assessment and job performance, but with the same prejudices operating in the measurement of both—assessment for the *status quo*. Perhaps the same people, unaware of selection issues, might be both carrying out the recruitment and working in the organization as managers making the job performance assessment. In such a situation, the organization is likely to select 'clones' and will probably find it difficult to adapt and develop in a changing world.

To avoid this, we must have very aware methods of job analysis and selection to work out the variety of ways in which jobs can be done well both now and in the future—and to select accordingly.

Another interesting consideration is that evolution itself provides a very effective selection model in which, as in the lottery, some randomness is introduced. Perhaps, then, we do need a mechanism to ensure that a diversity of people's abilities and working and thinking styles are represented within the organization for the 'learning organization' to become a reality.

The following contributions were received in writing after the meeting.

John F. Bell (*University of Cambridge Local Examination Syndicate*)

The suggestion that a graduated lottery be used for pass–fail decisions in examinations is both ill conceived and unnecessary. This can be demonstrated by using classical test theory. Let a candidate's test score X_i be partitioned as

$$X_i = \tau_i + e_i \tag{2}$$

where τ_i is the candidate's true score and e_i is a random error with mean 0 and variance σ^2. Let the pass mark be T, i.e. if $X_i \geqslant T$ then the candidate passes. This leads to four outcomes:

$$\tau_i \geqslant T \text{ and } \tau_i + e_i \geqslant T \qquad \text{(true positive);}$$

$$\tau_i \geqslant T \text{ and } \tau_i + e_i < T \qquad \text{(false negative);}$$

$$\tau_i < T \text{ and } \tau_i + e_i \geqslant T \qquad \text{(false positive);}$$

$$\tau_i < T \text{ and } \tau_i + e_i < T \qquad \text{(true negative).}$$

In effect, because no examination is perfect, the situation described in the paper applies because of the random error in the test.

The proposal is that something other than a simple cut-off value should be used. If it assumed that the probability of success given X_t, $P(s/X_t)$, increases as X_t increases, then it is obvious that the maximum expected number of successes for a given subset size is obtained by choosing the individuals with the highest values of X_t. The real problem with tests and examinations is to persuade selectors that statistics make the best predictions (e.g. Bell (1989), Dawes (1988) and Goldberg (1968)).

The effect of the system described in this paper depends on the importance attached to false positives and false negatives. Consider the example of an examination used to determine who should follow an expensive training course. The false negatives who would have a high probability of completing the course have obviously lost out on the benefits of the training course. The false positives have a greater probability of failing to make satisfactory progress on the course. This means that for the organization running the course the proportion of candidates satisfactorily completing the course decreases. This would mean that to obtain a particular number of successful trainees would result in increased provision and increased expense.

The only circumstance that a lottery should be used is to select a subset of *identically* qualified candidates. In practice, there is usually additional information that can be used to discriminate between candidates. The problem arises when this information is not relevant, e.g. 'hairism' and 'heightism'.

David Bellhouse (*University of Western Ontario, London*)

Many of the reasons stated in the paper for advocating random selection of people were earlier set out by Thomas Gataker, an English Puritan divine of the 17th century. In describing the division of property by lot, Gataker (1619), p. 165, says:

> 'True it is indeed that in the Civill Law all Appeale is denied ordinarily from the sentence of a Lot: But that is, not (as some of them fondly say) because the sentence of a Lot is the sentence of Fortune, or of God, who hath no superior in this world: but rather, as others, with better colour of reason; because this couse is taken for more speedie dispatch; because by flying from it in divers cases they shall but hinder either other from every comming to any issue; because a Lot is the most equall and indifferent course that can be and no corruption or partialitie can be charged upon it: and lastly because commonly it is by mutuall consent that matters are put thereunto, in which case their owne act justly concludeth either side.'

In fact, Gataker's entire book is devoted to a very reasonable discussion of the issues raised in Section 3.1 of the paper.

The quotation also points to two problems with the strategy advocated in the paper. The first is the law and the second is the consent of the public to participate in such a strategy. Should the strategy suggested in the paper become common, undoubtedly there will be legislation to regulate it. The law was and still is not keen on distribution by lot. It is avoided in all but exceptional circumstances. See, for example, the case of Knapton, Knapton *versus* Hindle and Others (All England Law Reports, 1941). Once over the legal hurdles there is the problem of acceptance by the general population. The problem is illustrated by the lotto or number lottery. One reason for the popularity of this lottery is the illusion of control that is present because players may choose their own numbers. I would expect that the general population would prefer to maintain any illusion of control in selection procedures rather than submit to a strictly aleatory selection. Consequently, the 'mutuall consent' among the general public would more probably be to accept a selection based on some kind of merit with admitted measurement error in conjunction with checks and balances provided by appeal procedures.

Nick Bingham (*Birkbeck College, London*)

This interesting paper is thought provoking and opens the field of discussion out well beyond the usual domain of Royal Statistical Society (RSS) discussions. In the same spirit, I offer a few thoughts of my own that were suggested by the paper, in random order.

(a) During the Vietnam war, the US draft was rightly seen as unfair, not because lotteries were not used—they were—but because being a student granted exemption. On the one hand, this meant that the burden of fighting the war fell on those already disadvantaged in society—the less well educated, the unskilled and, disproportionately, blacks. On the other hand, it led

within the universities to moral pressure on instructors (teachers)—which I remember well as I experienced it at the time—not to flunk (fail) students, as students dismissed from university on academic grounds immediately became liable for the draft. The distorting effect of the results of this moral pressure led to a wave of 'grade inflation', which persists in the system more than two decades later.

(b) Strict 'merit', though 'fair', can nevertheless be grossly distorting. In France, for example, with its Napoleonic heritage of meritocracy, the *Grandes Écoles* are the prestigious institutions, with mere universities seen as second best. The two-year intensive preparatory courses for the entrance examinations for the *Grandes Écoles* divert an immense amount of academic effort that could instead be used to revitalize the university system. In the UK, a similar distortion was found in the old Cambridge tripos system, which became more a 'sporting event' than genuinely useful academically. See for example G. H. Hardy's magisterial denunciation of the unreformed tripos in his presidential address to the Mathematical Association (Hardy, 1925).

(c) Membership of the Committee of Professors of Statistics (COPS) largely consists of white middle class males (such as myself)—perhaps the RSS should ask the COPS to break down its membership by gender, race etc. But would any of us advocate the use of lotteries here? Would we perceive unfairness in the *status quo*? As we are all statisticians, I suspect not—but I also suspect that the general public might.

(d) *Selection* is not the main issue in many cases, rather *what happens afterwards*—a matter of the *culture and practice* of the organization. To bring this home, we have only to recall the outcomes of a steady stream of settlements of successful lawsuits brought by women and members of ethnic minorities, about their treatment in the police and the armed forces—not to mention anecdotal evidence, in this and other areas.

Peter W. Donovan (*University of New South Wales, Sydney*)

The underlying problem is making discrete choices on the basis of numerical measurement of phenomena that can be measured at best crudely with numbers. I shall discuss four questions that emerge from this problem.

What is being measured?

What is being measured is of considerable interest to psychologists and economists. The intelligence quotient is something that can be measured and that appears to correlate strongly with academic ability as measured by examinations. This is discussed at length by Elo (1978). Measurement is complicated by the practical impossibility of running multiple repeats of tests to minimize the effects of short-term illness etc. The impression remains that such testing produces valid information and that it is possible to determine limits to its validity. Presumably Cambridge University determines grades of honours by requiring several different examiners to return marks on various projects and examinations. Such marks are then summed and various cut-off points are used to determine the grades. These grades have emerged somehow from subjective assessment with minor error quite possible. The 100 years since Edgeworth's time have not produced a better method. More comment on this topic is given by Beasley (1990).

Are the measurement standards constant?

Boyle does not consider whether the measurement standards are constant. To return to the example of Honours degrees, a choice between two people may involve comparing a degree from Cambridge University with another from Edinburgh University. In practice this will be done on the basis of partial information gathered a generation ago. In my experience there is a shocking lack of intelligently compiled comparative examination statistics for the Australian universities. Similar entrance examinations and modern computing technology should make such statistics feasible. I am sad to report that Elo's chess rating system has shown inflationary trends.

Can the handling of the situation be improved?

Any examination system such as that described, perhaps hypothetically, above for Cambridge University can issue its candidates with marks as well as grades. Thus a certificate stating that X received a mark of 799 in a system in which marks of 800–1000 yield first-class Honours and marks of 650–799 produce second-class Honours allows the graduate to say with some truth that if things had been a little different first-class Honours would have been granted. Fuller certification

reduces the problem. Edgeworth showed that there is a positive probability that a candidate with a total of 801 has achieved more than a candidate with 799. Any lottery in the grade determining process can only decrease the probability of justice being done.

Does it matter anyway?
At the time of writing I am awaiting my daughter's (discrete) result in the local equivalent of an eleven-plus examination. As the state has a continuous spectrum of secondary school opportunities I am not particularly worried. Would Conall Boyle be worried in these circumstances?

R. W. Farebrother (*Victoria University of Manchester*)
I should like to thank Mr Boyle for drawing Edgeworth's (1890) aleatory proposal to my attention. However, I think that Edgeworth would have preferred to reserve the word 'aleatory' for dice games and to have used a word based on the root *sors*, *sortis*, in the present context. This distinction will be familiar to many statisticians from the titles of Cardano's *Liber de Ludo Alea* and de Moivre's *De Mensura Sortis*; see David (1962).

It is interesting that Mr Boyle's scheme for allocating school places reverses the traditional relationship between lotteries and schools: instead of employing the holders of school places to draw lottery tickets, the school places themselves are to be awarded as prizes in a lottery.

Edgeworth and Boyle may both have underestimated the strong incentive to criminal activity when, as here, the outcome of a lottery significantly influences the quality of a person's own life or that of his children. In this context it is necessary to take measures to minimize the possibility of fraud. These include the following.

(a) If the draw is to be performed in public then it is sensible to require a large proportion of the participants or their agents to attend the draw.
(b) It is also sensible to employ a complete draw of all capsules, tokens or tickets to ensure that all candidates actually participate in the lottery to the stated extent.
(c) In certain circumstances it may be necessary to protect the identity of the prize-winners for a period of time by substituting code words for the names of the candidates.

These ideas are not new as all three precautions were features of the English state lottery of 1567.

The author's survey of Renaissance lotteries does not mention the method employed in the election of members of noble families to the Genoese senate. The game of chance developed from this source now forms the basis of the present British and many other national lotteries.

In this context, it may also be mentioned that the early history of casting lots is discussed in the early sections of David (1962), Pearson and Kendall (1970) and Kendall and Plackett (1977).

The awarding of entries in the *Encyclopedia Britannica* is a further example of a 'lottery'. Despite his weighty contributions to statistics and economics, F. Y. Edgeworth has not received an entry for more than 30 years, although his grandfather and aunt each have entries.

James Franklin (*University of New South Wales, Sydney*)
Everyone agrees that injustice arises from the variability of results about the true or correct outcome of a selection process. Boyle proposes a remedy that involves increasing that variability still further, by introducing a random element. The paradoxicality of a 'solution' that leads to a higher incidence of unjust outcomes is not squarely faced.

An increase in the average unfairness of outcomes might be outweighed by corresponding advantages. The advantage that Boyle suggests, a candid admission of the role of chance in selection processes, is perfectly real but is surely not the kind of benefit that could count substantially in comparison with the multiplication of injustices. It would appear also that implementing a system of lots as recommended would lead to some further inconveniences that are not mentioned by Boyle. For example, the population being selected from would change, as candidates with no chance of being selected on merit entered themselves in a lottery where they had nothing to lose.

It may be, however, that schemes of selection by lot merit attention in certain cases where the current selection process creates unsatisfactory biases in the population of candidates. It is arguable that the current system of selection for the Presidency of the USA makes election possible only for those who have acquired, one way or another, enough money to allow them to devote their lives solely to the attempt to become President. Selection by lot from all those who have shown minimal competence in

public office might be preferable. Another kind of example arises with examinations in some humanities subjects, where the variability between markers would be a scandal if published, not to mention the lack of correlation between the average marks awarded by politically correct examiners and actual merit. In such cases, marks by lot would introduce a welcome diversity into the pool of successful examinees; nevertheless, it is still difficult to believe that there would not be an even better outcome by moving to the method that Boyle so disparages, a traditional examination in classical Greek.

Harvey Goldstein (*Institute of Education, London*)

The problem of unreliability in selection is clearly important and Mr Boyle is to be congratulated for raising issues that are often dormant. The example of intelligence testing for school selection is interesting because the perception of 'fairness' here is so complicated. Not only is there the 'technical' issue of test reliability, there is also the issue of cultural bias, the 'self-fulfilling prophecy' aspect of judging effectiveness of a selection procedure by outcomes which may be influenced by the selection decisions, and the fact that the underlying rationale of a meritocratic procedure could be readily undermined by coaching for the test. The 'solution' adopted in this case was (in most cases) to abolish selection. This seems a perfectly rational response to the problem and is one that might be more widely considered. Mr Boyle alludes to this in his discussion of repeated lotteries and allowing repeated opportunities to individuals to minimize mistakes.

Mr Boyle claims several potential benefits for his procedures, such as less 'cramming', less cultural bias, a perceived greater fairness and a raised awareness of the unreliability of any given selection procedure. I am very much in favour of raising awareness of reliability issues, but I remain dubious about these other claims. If we set a minimum selection score above which candidates are automatically selected then many candidates' attentions inevitably will shift to this as a first-stage hurdle. Someone who just fails to reach it will have a non-zero probability of not being selected, whereas under the current common (single cut-off) system they would certainly have been selected. This could easily be seen to be 'unfair', since there will certainly be individuals, with lower true scores than this individual, who are selected. Cultural bias, if present in the test, will still be present whatever procedure is used and will contribute, if recognized, to perceptions of unfairness. Will cramming surely now take place to achieve the minimum selection score?

John Haigh (*University of Sussex, Brighton*)

Those affected by decisions are more likely to accept them if they can have confidence in the methods by which these decisions are reached. Taking the problem of Section 4.3, suppose that the aim is to select 25% of a group of 1000 pupils to attend grammar schools, on the basis of intelligence quotient (IQ) tests. Simulation can be useful to explore the consequences of different methods of reaching a decision. A random sample of size 1000 from an $N(100, 15^2)$ distribution leads to values $\mu(1) < \mu(2) < \ldots < \mu(1000)$ which represent the 'true' IQs of the population. The *score* $X(i)$ of the ith pupil's IQ test might reasonably be assumed to follow an $N\{\mu(i), \sigma^2\}$ distribution, for some unknown variance σ^2, which indicates the *repeatability* of these tests. (Vernon's estimate that 10% of candidates who pass the test would fail it on a second attempt is consistent with σ being approximately 4. Of course low values of σ are desirable.) The decisions are to be made on the basis of the scores $\{X(1), X(2), \ldots, X(1000)\}$.

Without using a lottery, the 250 pupils with the highest scores will be selected. Random chance will lead to some pupils with high intrinsic IQs not chosen, and others with lower values of $\mu(i)$ leaping ahead. This might be expected to occur more often with Mr Boyle's lottery cascade: but how much more often? It is interesting to compare the outcomes of the two methods for different values of σ. Table 4 shows the average outcomes, for one set of simulation runs, although other runs gave very similar output.

The data emphasize the importance of σ being small. Table 4 suggests that, so long as σ does not exceed 4, very few injustices will arise when taking the $X(i)$ as firm indicators. But, when $\sigma = 8$, even someone in the top 50 fails, and a few with below average IQ would pass.

The outputs for the lottery method are fairly similar across the three values of σ, for pupils ranked 551–950. However, even when $\sigma = 2$, not all those in the top 50 will be selected. (The same also occurred in a run with $\sigma = 1$.) When $\sigma = 4$, 30 of those in the top 150 would fail. Would such outcomes be seen as acceptable by the public?

Table 4. Numbers of pupils selected from each band, for different values of σ†

True rank, using the unknown $\{\mu(i)\}$	*Numbers selected, for the following values of* σ*:*					
	$\sigma = 2$		$\sigma = 4$		$\sigma = 8$	
	Order	*Lottery*	*Order*	*Lottery*	*Order*	*Lottery*
Top 50, $\mu > 125$	50	48	50	48	49	44
Next 100, $116 < \mu < 125$	100	74	99	72	83	68
Next 100, $111 < \mu < 116$	81	54	72	54	57	49
Next 100, $106 < \mu < 111$	19	42	24	40	34	39
Next 100, $102 < \mu < 106$	0	26	5	24	15	23
Bottom 550, $\mu < 102$	0	6	0	12	12	27

†The 'order' column refers to using the strict order of the values $X(i)$; the 'lottery' column is using the lottery cascade on the same data.

Paul March (*LandMark Insurance, London*)

Segmented group

The paper considered only the situation where the applicants with the highest scores were selected. Some organizations (e.g. a mixed ability school) might try to achieve a 'mix' of applicants by banding the results and selecting the best from each band rather than the group as a whole.

The difference here is that you are chosen on the basis of being below and closer to the top of a band than your peers are. In this case an error of increasing your score one point above the top of a band would virtually exclude you as you would be the last in the next higher band. This would lead to marked differences in consequences for a very small movement in score.

Constraining successful outcomes—it is a lottery now

No matter how fair (without any irrelevant factors or any bias etc.) or accurate (true predictive measure) any assessment is there will always be a lottery element. Where there is a fixed number of successful outcomes greater than the number of worthwhile applicants then there will be a lottery element. The decision will be based on the relative placing with the other applicants. As the relative strength of the competition will vary from application to application (possibly greatly; possibly very little) there will be this lottery element. It was not a true lottery but a type of graduated lottery. The paper is therefore changing the lottery element rather than introducing one.

Non-constraining successful outcomes

Consider where the number of successful outcomes is only limited by the number of applicants (as in the case of an examining body) and is independent of the mix of applicants. The lottery element due to the effect of the competition is zero. This type of competition is different from the type addressed in the paper and Mr Boyle is not recommending his lottery strategy for this non-constraining scenario generally but only for some type of borderline case between the grades.

Lottery and the Bible

The use of lots and the seeking of guidance in the *Bible* is a very wide subject. The lots could fall on a particular outcome (in a similar way to a lottery) but could also not give any answer at all which is very different from a pure lottery. Really, this aspect of the subject needs a whole paper to describe the use of lots in the *Bible* adequately.

The **author** replied later, in writing, as follows.

The proposal to introduce lotteries as part of the selection process draws on the technical insights of statistics, as well as psychology, and is intended as a practical administrative technique. It is not surprising therefore that these various requirements may sometimes clash. These I shall try to explain in response to the comments.

Many commentators have made broadly similar points, so I shall group them together.

Error and uncertainty

Adding a lottery to a lottery must increase error and uncertainty is a comment from Bartholemew, Evans, Bell, Franklin, Haigh, March and Donovan, in various forms. We are all agreed that the tests which seek to discriminate between candidates are subject to variability, that at best they constitute a graduated lottery. I agree that tampering with test scores by adding a further process of randomization can only make the results less reliable. Whatever other benefits a lottery might bring, systematically less reliable tests would not be a fair way of dealing with the candidates. It would also seem to throw into doubt Edgeworth's claim that adding a lottery stage to civil service examinations would *ex hypothesi* still produce the same proportion of good candidates.

I would not of course advocate tampering with test scores; they should be issued as found, ideally with a warning about expected unreliability. It is at the next stage of the process, when selection from a short list takes place, that I advocate the use of a lottery. So why should we not accept the test score and appoint or award the top scorer?

Firstly, consider non-linearity: although it is true that ability may be related to test scores, above a certain threshold higher scores may not indicate any higher potential. Evidence for this can be found in Eysenck (1962) in a graphic (p. 26) which shows how success as trainee officers and pilots increased with the intelligence quotient (IQ) score. This is not a linear relationship; beyond a score of 120 there is hardly any improvement in performance. More evidence of this 'plateau' effect is given in Blum (1978) (p. 81) who quotes Frank Barron

> 'that for certain intrinsically creative activities a specific minimum IQ is probably necessary to engage in the activity at all, but beyond the minimum, which is often surprisingly low, creativity has little correlation with scores on IQ tests'.

Secondly, test scores are unreliable indicators of a limited repertoire of abilities. Invariably many crucial attributes are not captured by the test. Dr Deming was fond of quoting 'The most important figures for management of any organization are unknown and unknowable' (Neave (1990), p. 151). The test score plus a lottery will still tend to select those with higher ability as indicated by the test. But it will also allow through a sprinkling of all the other talents which may have been systematically excluded by the test.

A further problem noted by Creighton, who is involved in selection tests for today's civil servants, is the danger of cloning, i.e. appointing candidates from very similar backgrounds and with very similar mind sets. By introducing some variety, a lottery could act like the grit in the oyster, and produce a few pearls who could greatly benefit the organization.

It was perhaps a combination of these features which Edgeworth had in mind when he declared that the same number of good candidates would get through.

Efficiency

Hawkins, Kempton, Bingham and Goldstein comment that a lottery would not be as efficient. Efficiency should mean reducing the total effort of both the organization and the pool of candidates while still producing appointees who are just as good. As a result of a lottery stage, effort by the candidates would be greatly reduced. For the organization the effort saved in the selection stage would not be great, and might be outweighed by the cost of appointing or accepting a dud. Hiring a worker or admitting a student to a lengthy course commits the organization to considerable outlay. But, even for the organization there is some compensation.

Firstly, following the argument of the previous paragraph, although a few more duds would get through, so would a number of pearls. This should mean a better overall result for the organization.

Secondly, an organization which uses an elaborate selection process may be reluctant to admit that they had failed, and throw good money after bad in trying to rescue themselves from a poor decision. An organization using a light test plus a lottery might find it easier to remove non-performers.

Thirdly, coping with claims for unfair dismissal and discriminatory employment practices is creating a major burden for organizations. It was reported in the *Daily Mail* on August 8th, 1997, that Lewisham Council, London, requires a 40-page document to describe its equal opportunity policy. Dealing with court judgments on unfair employment practices places great burdens on organizations and is a major distraction from their main activities. A lottery as part of the selection process ensures greater fairness and could be an effective defence against claims, helping to avoid costly and time-consuming litigation.

Tests—improve or abandon

Comments on validity of tests came from Lewis, Derbyshire, Goldstein and Bellhouse, to the effect

that if we are to have tests then they should be sound, and we should try to improve them. This involves probing the validity of current tests and putting resources into developing better tests. Certain moral hazards appear: tests are commercial products, and there is always a temptation to oversell them, and for organizational bureaucrats to act defensively by not questioning their validity. We should be concerned that the addition of a lottery to the process is not used to weaken the drive towards improving the validity of tests.

A different problem of validity is suggested by Hawkins, Goodwin and Creighton—the inevitability of cultural bias in any test however refined. It is always useful to expose and eliminate such bias in the tests, but should we go further, adding an 'appraisal' stage, or affirmative action? I think not. Both are suspect—appraisal, because it is far less reliable as Vernon (1957) explained. Affirmative action is intrinsically illiberal, and it is criticized because it helps those most who need it least. Far better I would contend would be the 'light test' advocated by Edgeworth, which should be far less daunting to 'outsiders'. The subsequent lottery would not eliminate cultural bias, but it would significantly alleviate it. Crucially the lottery stage would act against all forms of bias, not just those that are recognized. It would also deal with the problem raised by Derbyshire that an underachieving group may, through a change in circumstances, start to overperform: witness the startling improvement in girls' performance in General Certificate of Secondary Education examinations. A lottery is a much more flexible and responsive method for dealing with *all* forms of cultural bias in tests.

Outer gate-keeping—who should be let into the lottery?

Comments came from Goodwin, Carr-Hill, Kempton and Franklin on whether the selection by lottery process ought to be extended wider. The lottery scheme that I described is, in essence, a method of choosing winners from a short list. This involves two total exclusions: those scoring below a threshold on a test *and* those who are not even allowed to enter the test. As to the first, a cut-off value is theoretically unfair as Goodwin says, because there is always a diminishingly small probability that a very low scorer *might* have a much higher true result. My reasons for not proposing a graduated lottery involving all the candidates is firstly administrative, to simplify the process. But we should also distinguish between the theoretical long tail of the normal distribution and what I believe is the practical reality that a definite cut-off exists at some middling value.

A problem which I have evaded was what could be called 'outer gate-keeping': who should be allowed in for consideration. A practical example of this is the waiting list procedure for social rented housing (council houses in UK parlance). Even to be put on the list you must exhibit residency and other qualifications. I am encouraged by the comment from Kempton about success in selecting seed varieties—adopt an 'open-door' policy at first, allowing in all comers, but subject them to an initial lottery to reduce the numbers to be tested. However, this idea might create many more problems, and perhaps it is best to leave the outer gate-keeping problem to administrative procedures, which should be overt and challengeable.

Fairness by lottery—people would not understand, and they do not want it anyway

Comments by Hawkins, Bartholemew, Franklin, Goldstein, Haigh and Bellhouse relate to what could be called the psychology of fairness. What is perceived as fair is often highly subjective; what may be seen as fair to one is not seen that way by another. We are dealing with not just peoples' understanding but also their feelings. However cleverly we explain that a lottery is technically fair, what if people do not understand or, worse, dimly understand but have a feeling that they do not like it? This seems to be the case with the present proposal. Elster (1989) as I noted in the paper thinks that it is better to let people labour under a delusion, if only to satisfy some primitive urge to find a human scapegoat for an allegedly unfair decision. This I believe is wrong: responsible professionals have a duty to alert their public when something is wrong, even if the public is comfortable in its delusion. A lottery would have the benefit of reminding people of the uncomfortable reality that many decisions are no more than a toss-up.

But is the general public so ignorant about matters probabilistic? It may be true that surveys show widespread inability to calculate odds, or to explain relative risk factors. But this does not preclude an *intuitive* grasp of probability. Why else would children pick teams by using counting games, or their elders start a game with the toss of a coin? Historically, the ancient Greeks and Renaissance Italians insisted on the lottery to choose their leaders. This was well before a theory of probability had been developed, so again it must have been an intuitive understanding. Perhaps the citizens of Athens and Florence were indeed more intelligent than, say, present-day Englishmen (a view held by Sir Francis Galton!). If so it is a reflection on our lack of success in educating the young.

Rather than expecting members of the public to clamour for change, I guess that pressure to adopt lotteries in the process of selection would arise in the knotty field of equal opportunities. This would become more acute when, as is proposed by communitarians, a high degree of responsibility is devolved down to small local units. The sophisticated selection procedures which are required for fairness would be difficult for these community-based organizations to sustain, so the quick, cheap and fair lottery process would be a life-line.

Other points in brief

Bellhouse suggests that a lottery proposal would run foul of the law. How legal processes involve fairness or equity is a fascinating question but goes beyond this paper. On the specifics of lotteries used for selection, Elster (1989) quoted several legal cases which tested and accepted the idea, so perhaps the law is not quite so hostile.

Carr-Hill and Donovan give examples which also do not fall within the ambit of this paper: what if there is only one candidate, or if all candidates are allocated to one of a range of options? Both could involve unfairness, but neither has the element of scarcity.

I conclude with two suggestions which at first glance might be thought whimsical: from Goodhart that research money be handed out by lottery to academic institutions, and by Evans that the members of the House of Lords be chosen in this way. On further reflection, both of these proposals are eminently sensible! Choosing a political assembly by lot would begin to establish democracy as the ancient Athenians understood it. The idea has been revived by Burnheim (1985) under the title 'demarchy', but this goes far beyond my own modest proposal which relates only to organizations selecting people.

References in the discussion

All England Law Reports (1941) Knapton, Knapton *v.* Hindle and Others. In *All England Law Reports*, vol. 2, pp. 573–576.

Beasley, J. (1990) *The Mathematics of Games*. Oxford: Oxford University Press.

Bell, J. F. (1989) Well I know someone who . . . or the curse of the concrete. *Professl Statistn*, **8**, 11–13.

Blum, J. M. (1978) *Pseudoscience and Mental Ability: the Origins and Fallacies of the IQ Controversy*. New York: Monthly Review Press.

Burnheim, J. (1985) *Is Democracy Possible?: the Alternative to Electoral Politics*. Cambridge: Polity.

Carr-Hill, R. A. (1989) Assumptions of the QALY Procedure. *Socl Sci. Med.*, **29**, 469–477.

———(1991) Health related quality of life measurement—Euro style. *Hlth Poly*, **20**, 321–328.

Carr-Hill, R. A. and Morris, J. (1991) Validating the 'Q' in QALYs: a cautionary note. *Br. Med. J.*, **303**, 699–701.

Cox, D. R., Fitzpatrick, R., Fletcher, A. E., Gore, S. M., Spiegelhalter, D. J. and Jones, D. J. (1992) Quality-of-life assessment: can we keep it simple (with discussion)? *J. R. Statist. Soc.* A, **155**, 353–393.

Cresswell, M. J. (1996) Defining, setting and maintaining standards in curriculum-embedded examinations: judgemental and statistical approaches. In *Assessment: Problems, Developments and Statistical Issues* (eds H. Goldstein and T. Lewis), pp. 57–84. Chichester: Wiley.

David, F. N. (1962) *Games, Gods and Gambling*. London: Griffin.

Dawes, R. M. (1988) *Rational Choice in an Uncertain World*. Orlando: Harcourt Brace Jovanovich.

Edgeworth, F. Y. (1890) The element of chance in competitive examinations. *J. R. Statist. Soc.*, **53**, 644–663.

Elo, A. (1978) *The Rating of Chessplayers, Past and Present*. London: Batsford.

Elster, J. (1989) *Solomonic Judgements*. Cambridge: Cambridge University Press.

———(1992) *Local Justice: how Institutions Allocate Scarce Goods and Necessary Burdens*. Cambridge: Cambridge University Press.

EuroQol Group (1990) EuroQol: a new facility for the measurement of health related quality of life. *Hlth Poly*, **16**, 199–208.

Eysenck, H. J. (1962) *Know Your Own I.Q.* Harmondsworth: Penguin.

Finney, D. J. (1958) Statistical problems of plant selection. *Bull. Int. Statist. Inst.*, **36**, 242–268.

Gataker, T. (1619) *Of the Nature and Use of Lots*. London: Griffin.

Goldberg, L. R. (1968) Simple models or simple processes?: some research on clinical judgements. *Am. Psychol.*, **23**, 483–496.

Goodwin, B. (1992) *Justice by Lottery*. Hemel Hempstead: Harvester Wheatsheaf.

Greaney, V. and Kellaghan, T. (1996) The integrity of public examinations in developing countries. In *Assessment: Problems, Developments and Statistical Issues* (eds H. Goldstein and T. Lewis), pp. 167–188. Chichester: Wiley.

Hardy, G. H. (1925) The case against the Mathematical Tripos. *Math. Gaz.*, **12**, 309–316.

Hunt, S. M. and McKenna, S. P. (1991) Letters. *Br. Med. J.*, **305**, 645–646.

———(1993) Letters. *Br. Med. J.*, **307**, 125–127.

Kempton, R. A. and Talbot, M. (1988) The development of new crop varieties. *J. R. Statist. Soc.* A, **151**, 327–341.

Kendall, M. G. and Plackett, R. L. (1977) *Studies in the History of Statistics and Probability*, vol. II. High Wycombe: Griffin.
Lancashire County Council (1997) Agenda for the meeting of the Education Equal Opportunities Sub-Committee held on July 1st, 1997.
Martin, S., Rice, N. and Smith, P. (1997) Risk and the GP budget holder. *Discussion Paper 153*. University of York, York.
Neave, H. (1990) *The Deming Dimension*. Knoxville: SPC.
Pearson, E. S. and Kendall, M. G. (1970) *Studies in the History of Statistics and Probability*, vol. I. London: Griffin.
Vernon, P. E. (ed.) (1957) *Secondary School Selection—a British Psychological Society Inquiry*. London: Methuen.
Vincent, R. (1996) Assessment in the workplace. In *Assessment: Problems, Developments and Statistical Issues* (eds H. Goldstein and T. Lewis), pp. 231–244. Chichester: Wiley.
Ware, J. E., Snow, K. K., Kosinski, M. *et al.* (1993) *SF-36 Health Survey Manual and Interpretation Guide*. Boston: New England Medical Centre.
West, P. (1987) The value of health economics. *Rad. Commty Med.*, **24**.
Williams, B. (1981) *Moral Luck*. Cambridge: Cambridge University Press.