STATISTICAL AND DECISION THEORETIC ASPECTS OF EXAMINATION ASSESSMENT

FRENCH, S.
Statistical Laboratory
Mathematics Department
University of Manchester
Manchester M13, 9PL

ABSTRACT

At a recent conference in Innsbruck, held in memory of Bruno De Finetti, attention focused on one of his aphorisms: analysts «should think about things». This paper seeks to do precisely that in the context of public examinations in England and Wales. It attempts to think about those quantitative things that are done to marks in the process of assessing candidates scripts. Public examinations are central to our education system: hundreds of thousands of candidates enter them each year. In assessing the candidates performances, the examination boards must absorb, analyse and combine literally millions of marks. To do so they must use many quantitative procedures. What is the validity of these procedures? What is their purpose and do they have their desired effect?

Key words: appropriate statistics; descriptive and normative modelling; educational assessment; exchangeability; latent traits; meaningfulness; measurement theory; multi-attribute value theory; parametric and nonparametric statistics; scale types.

1. INTRODUCTION

In common with other Bayesians, I pride myself on a coherent approach to analysis in both the technical and everyday sense of coherence. Yet it is all too easy to confine attention to the coherent analysis of a model without considering the implications of the Bayesian philosophy for the generation of the model itself and, indeed, for the conception of the context in which the model is developed and analysed.
The essence of the Bayesian approach, it seems to me, is the recognition that, as individuals, we must continually express beliefs, preferences, judgements, etc., and act according to these. In doing so, we wish to be as rational, as consistent and, perhaps, as fair as possible; and we seek ways of thinking which help us achieve these ends. A Bayesian is not simply someone who updates a prior by a likelihood and then maximises an expected utility. He is someone who thinks carefully about how to encourage—he hopes ensure—consistency and coherence in his perceptions, judgements and actions.

In French (1986, Chapter 9) I explore some implications of this view for decision analysis: here I wish to explore some of its implications for the way in which we conceive of certain aspects of the English and Welsh public examination system. Such discussion is timely, because our system is currently undergoing much reappraisal and restructuring. However, I believe that the import of this discussion is not confined to our system: it is relevant to debates on examinations elsewhere. Moreover, I hope that the paper contributes to the general literature of mathematical, statistical and judgemental modelling.

Public examinations (i.e. examinations set and assessed by boards which are at least partially independent of the candidates, schools, colleges, etc.) play a central role in our education system. The majority of school-leavers take some form of public examination designed to certify academic, technological, trade and craft skills. The current restructuring means that it is likely that in future an even higher proportion of school children will be assessed, and, moreover, they will be assessed more often and in more subjects that at present. The system is already complex and about to become more complex. I shall avoid some of the complexity, however, by focusing on the English and Welsh Advanced Level General Certificate of Education (GCE A levels), taken mainly by 18 year old school-leavers. The Ordinary Level GCE has just been combined with the Certificate of Secondary Education (CSE) to form the General Certificate of Secondary Education (GCSE), which will be taken mainly by 16 year olds. In time the GCSE is expected to become far more explicit criterion referenced than the GCE O and A levels and the CSE have been; but at present, apart from a much greater emphasis on continuous and other forms of internal assessment, it differs little in fundamental form from GCE A level.

GCE A levels are subject-based. There is no requirement to pass or
matriculate in groups of subject. Candidates are awarded a separate grade for each subject taken, and the choice of subject is left to the candidates in consultation with their teachers and parents. The form of examination, naturally, varies from subject to subject. Apart from sitting formal written examination papers, candidates may be required to submit coursework or projects, or to be assessed practically or orally. Within the formal papers they may be required to write essays, answer short, structured questions or multiple choice items. In reporting the candidates performances the examination boards have to draw together information gathered by these various assessment instruments. They must ensure that all work is marked fairly and consistently with due credit being given for good work. They also seek to award final subject grades which are sufficiently detached from the particular teaching, syllabus, examination, examiner, etc., so as to have meaning to those receiving and using them.

The entry to a single subject may be several thousands, perhaps tens of thousands. Thus the tasks facing the boards, difficult enough in themselves, are awesome in their magnitude. Certainly, it is impossible to assess each candidate's performance as individually and holistically as some might wish. At some point in the assessment process, usually early on, each candidate's performance is summarised as a set—or profile—of marks. Thereafter, the assessment is based almost entirely upon this profile. Marks are numbers and therein lies a danger. Any set of numbers can be added, averaged, compared, statistically analysed, etc., even though for a particular set in may be quite meaningless to do any of these. The procedures used by examination boards do add, average, compare and analyse marks. Do they do so in a valid, meaningful way?

Comprehensive descriptions of the English and Welsh public examining system may be found in the books by Christie and Forrest (1981) and Mathews (1985). French et al (1986a, b) contain a brief description of the various procedures involved in assessing an archetypal examination.

The plan of the paper is as follows. In the next section I suggest—somewhat contentiously but, nevertheless, briefly—the purpose of a public examination. Then I note the distinction between descriptive and normative measurement, and indicate how both types of measurement interact in the procedures involved in public examining. Subsequent sections deal with more technical topics: the theory of scale types and their implications for statistical inference; exploratory data
analyses; latent trait models and their interpretation; and the use of multi-attribute value theory to construct appropriate mark aggregation functions.

Many of my remarks and arguments derive from joint work with John Slater, Marilena Vassiloglou and Alan Willmott. Indeed, any sanity detectable in the following is probably due to their influence: the insanities remain my responsibility. I would also absolve from any complicity the several examination boards with which I have connections: I do not purport to represent their views.

2. THE PURPOSE OF PUBLIC EXAMINATIONS

I lack the space to discuss fully the purpose of public examinations or, more precisely, that of GCE A level. So here I simply summarise arguments that may be found elsewhere (French et al., 1986a, b). However, in summarising that discussion I do not wish to underplay its importance. It defines the context of all that is to follow and thus is central to my later arguments: to discuss the validity and appropriateness of procedures one must be clear about their purpose.

At a naive level, the purpose of GCE A level examinations is simply to certify the achievements of the candidates in particular subjects. However, the question that concerns us is what does «certify the achievements» mean. The current view in the educational assessment literature seems to have been developed, by and large, from the psychometric concepts and theories used to analyse psychometric tests, i.e. IQ tests, etc. These theories have at their base a belief that (a) inside each person lies something that might be called his or her «ability or level of achievement» in the tasks elicited by the tests, (b) this entity can be quantified on an objective, unidimensional scale, and (c) the purpose of the test is to gather data from which it can be estimated. Educational assessment theories as applied to public examinations certainly recognise that they differ greatly from psychometric tests, especially in their complexity and breadth. Nevertheless, there is still the feeling that the purpose of examinations is to gather data from which some «level of achievement in the subject examined» may be estimated. In this sense, theories of educational assessment and psychometric theories differ little.

I should, perhaps, qualify the previous paragraph with the remark that the vehemence with which this belief is stated varies greatly from
particular theory to particular theory. Many authors adopt this belief purely pragmatically. For example, the following quotation is typical of many of the writings in this area:

«However, to admit “true ability” into the discussion opens up some considerable difficulties, not the least of which is that although the discussion only makes complete sense if the ability in question is essentially one dimensional, this does not seem likely to be the case in any context of real interest. But aside from using some device such as reporting “profiles” for each candidate there seems to be no way of dealing with the problem, and pragmatically one has to interpret the overall ability being summarized in a final (single) mark as some combination of the various component abilities, including such things as examination technique, amount of preparation and so on.»

(Biggins et al., 1986, pp. 150-151)

Pragmatic though this view is, and it is certainly softer than the view of pure psychometric theory that I paraphrased above, it has at its core the belief that there is something inside the candidate that is being measured and reported.

I, and others, hold a very different view. The purpose of a public examination is not to measure in some objective sense something directly about the candidate. I do not believe that there is any entity withing the candidate that could be measured in such a way. Instead the purpose is to report the judgements of examiners. Moreover, the examiners do not make their judgements about some trait that they postulate to exist withing the candidate, his or her ability, achievement, or whatever; but rather they are concerned with the quality of performance, as they perceive it, within the candidate’s scripts and, in the case of practicals, coursework and orals, within the processes observed during the assessment. Public examinations are a means of reporting a detached academic judgement of the work of a candidate. The grades or marks do not encode a measurement of an entity within the candidate; they encode the examiner’s judgement of the quality of the candidate’s performance.

This distinction is not just an irrelevant and esoteric academic point: it has significant implications for the form of many of the numerical procedures used to manipulate marks in the assessment process. These procedures are not there to estimate candidates’ abilities in the presence
of «measurement errors». Nor, as I shall argue in the next section, do they have the purpose of estimating the examiners' judgements. Their purpose is more subtle. It is to help the examiners' form their judgements in a manner that is fair and consistent to the candidates. It is to arguing this that much of the next section is devoted.

3. THE DISTINCTION BETWEEN DESCRIPTIVE AND NORMATIVE ANALYSES

The distinction between descriptive and normative analyses is one with which Bayesians are well familiar. All too often we are pilloried because Bayes' Theorem does not model satisfactorily the heuristic rules used by individuals to update their beliefs in the light of data. Ah yes, we respond, but Bayes' Theorem does not describe how we do update our beliefs; it prescribes how we should update our beliefs if we wish to be rational in a certain sense. However, familiar as we may be with this distinction, we do not often explore it further and realise that it has implications for the modelling of other of our judgements than beliefs and preferences. (See, e.g., Dempster, 1985; French, 1986.) So what is the distinction?

The purpose of descriptive analyses is to investigate how things are. The models built are there simply to depict the system under observation. There is no intention that the analysis of itself will change the system in any way. True, Heinsenberg has shown that ultimately there are limits to an observer's ability not to interfere, not to affect the system he observes; but the intention in a descriptive analysis must be to interfere as little as possible. Certainly, there should never be an intention to change the system through its observation.

The purpose of normative analyses is quite different. They are intended to change, make consistent and, indeed, construct an individual's beliefs, preferences, values, etc. Numbers and models are used to help an individual think about things in a very active way. The individual is asked for his subjective feelings in the area of concern. His answers are encoded and analysed against certain axioms which he has previously agreed are canons of rationality. Points of consistency and inconsistency between his judgements and the axioms are identified and exhibited to him, thus enhancing his understanding of himself and, more
importantly, implicitly guiding the evolution of his beliefs, preferences and values in the direction of consistency.

The distinction that I am drawing here is seldom so clear in practice. Many analyses have both descriptive and normative aspects. None the less, for any analysis it is important to distinguish which are which. Calculations and interpretations of models can only be meaningful and valid if the purposes to which they are to be put are clearly understood.

Many of the ways in which numbers are used in examination assessment are predominantly normative, far more so than has been appreciated in the past. Dwelling on one such may make clear exactly what I mean by a normative analysis. Consider how an examiner may mark a batch of scripts. For convenience, we shall assume that he is the sole examiner and not part of a team marking the same paper.

Before he begins, he will have drawn up a rough marking scheme to reflect the qualities of work for which he is looking. Suppose then that he begins marking. Candidate A’s script is assessed, then B’s, etc. As he proceeds, he may refine the marking scheme because he becomes aware of ambiguities within it. However, such changes are not ones upon which I wish to focus. Rather, I want to consider the following not uncommon situation. Suppose that at some point the examiner becomes uncomfortable with the results of his marking. Some candidates seem to be accumulating more marks than he would judge holistically to be their due, and others correspondingly less. He will pause and reflect upon this and —categorising very broadly— may come to one of two conclusions. Either the marking scheme that he has constructed does not truly represent his judgements of quality of work or his holistic judgements are false because, perhaps, they overlook some aspects of candidates’ work which he has rewarded and, indeed, intended to reward in his marking scheme, but which he has forgotten in judging the scripts holistically. In the first case he will revise his marking scheme accordingly and in the second he will modify the processes which he uses in judging scripts holistically. In both cases an inconsistency between a supposed numerical representation of his judgements and his judgements themselves has caused him to pause, reflect and revise the processes that he is using. His judgements of quality of work are evolving: the marking procedure is guiding them towards consistency.

Several of the procedures used in examination assessment have a similar objective of guiding the examiners’ judgements towards consis-
tency and overall fairness to the candidates. The numerical aspects of these procedures are, therefore, predominantly normative. As indicated above, the design of marking schemes and the process of marking involve many normative procedures. The procedures used to aggregate marks across an examination's components are also normative, although this has seldom been recognised. Their purpose is to help examiners make an overall assessment of each candidate's scripts in the light of the importance of the various components. Consider.

A candidate has produced several pieces of work in an examination. He has answered question papers, done practicals, been examined orally, etc. These have been marked and the examiner now has to give an overall assessment. Many authors have suggested that the examiner faces a problem of estimating the candidate's overall ability and hence devised aggregation rules based upon multivariate statistical theory (French, 1985). But in the previous section I suggested that there was no such entity, in an objective sense, as the «candidate’s overall ability». There is only the examiner's judgement of the overall quality of work exhibited in the candidate's scripts. The process of designing and applying the aggregation rule is a normative one which, has the primary purpose of helping the examiner form his judgement in as fair and consistent manner as possible. We shall discuss this further in a later section.

Not all the procedures used in examining are normative: many are descriptive. For instance, public examinations generally have very many candidates. Thus it is impossible for a single examiner to mark all the scripts; teams of examiners must be used. This immediately introduces the problem of coordination. How can the chief examiner be sure that his team are marking «as one»: i.e. would they make the same judgements on the same scripts? Undoubtedly, the most important procedures in coordination involve interaction between the examiners: meetings, discussion, doublemarking of example scripts, etc. None the less, some descriptive numerical analyses are carried out. The chief examiner may compare the mark distributions of his assistant examiners to see if they are similar and, hence, under the assumption that each marked a similar population of candidates, whether the team are marking to a common standard. In the next two sections I discuss some of the statistical procedures that may be used to help in such comparison.

Another area in which descriptive analyses are important is the
following. In order to judge the quality of a candidate's work, particularly the overall quality, the examiners need to be fully aware of the context in which it was elicited. They need to have an idea of the «difficulty» of their question papers. Much of their understanding will come from their involvement in the design of the question papers, but not all of it. Candidates often read a question differently than was intended and hence find it «harder». The examiners will become aware of this in part from their reading of scripts, but statistical analyses can support their impressions, and occasionally suggest them to be suspect. Again we shall discuss this further in the following.

4. THE THEORY OF SCALE TYPES AND THEIR IMPLICATIONS FOR STATISTICAL INFERENCE

In both descriptive and normative analyses numbers are used to represent various entities and relations between them. I have already cautioned that any set of numbers can be added, averaged, compared, statistically analysed, etc., whether or not doing so is meaningful for that set. My purpose in this section is to make this warning more precise. It has implications both for the conduct of quantitative analyses within examination assessment and for the appropriateness of parametric and nonparametric statistical methods.

Firstly, what do we do when we build a quantitative model? How are numbers used? Essentially, quantitative models replace qualitative relations such as

i) «this stick is longer than that stick», or
ii) «I prefer object A to object B»

with quantitative ones. Thus i) would be modelled by associating a larger number, its length, with the first stick than with the second; and ii) by associating a larger number, my utility, with object A than with object B. The reasons for replacing «words with numbers» are manifold. Quantitative models are more concise and allow a greater precision of statement. It is easier to calculate with numbers than to derive results with words. By expressing everything within a common quantitative language, we are more able to detect common patterns. And so on. Of course, we cannot use numbers to represent qualitative relations at whim, nor can we deduce that some qualitative relation exists from each
and every pattern we see among the numbers. We must be careful that all the quantitative relations that we use in our analysis represent qualitative relations that truly exist and we must guard against being misled by extraneous quantitative relations that are simply artefacts of any arbitrary choices in the numbers used.

A simple example will give these points some substance. Consider three candidates who have taken an examination with three papers. Table 4.1 gives a set of possible marks: assume that each paper has been marked with a possible maximum of 100.

<table>
<thead>
<tr>
<th>Paper</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate A</td>
<td>63</td>
<td>59</td>
<td>66</td>
<td>188</td>
</tr>
<tr>
<td>Candidate B</td>
<td>65</td>
<td>57</td>
<td>69</td>
<td>191</td>
</tr>
<tr>
<td>Candidate C</td>
<td>69</td>
<td>56</td>
<td>68</td>
<td>193</td>
</tr>
</tbody>
</table>

The first instinct of examiners is to compare the totals. For instance, candidate B’s total in higher than A’s. So it appears that B did better overall. But is this conclusion justified? The marks on the papers represent the examiners’ judgements of the quality of the candidates’ work on each of the papers. Thus the quantitative relation between 65 and 63 on Paper I represents the ordinal judgement that B’s work on paper I was of a higher quality than A’s. Similar comments apply to the Paper II and III marks. However, it does not necessarily follow that the ordering of the total marks represents some judgement on the relative overall quality of their work.

Suppose that three different examiners have marked the three papers. Suppose also that the examiners have only made ordinal judgements. They have judged the relative ordering of the candidates’ work, but have made no judgements of the «amount» by which the quality of one candidate’s work exceeds that of others. (We shall make this statement more precise in section 7: for the present a vague state-
(ment will suffice.) The marks in Table 4.2 would also represent their judgements.

<table>
<thead>
<tr>
<th></th>
<th>Paper</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
<td></td>
</tr>
<tr>
<td>Candidate A</td>
<td>59</td>
<td>65</td>
<td>46</td>
<td>170</td>
</tr>
<tr>
<td>Candidate B</td>
<td>60</td>
<td>61</td>
<td>48</td>
<td>169</td>
</tr>
<tr>
<td>Candidate C</td>
<td>61</td>
<td>58</td>
<td>47</td>
<td>166</td>
</tr>
</tbody>
</table>

In Table 4.2 the ordering of the totals reverses that in Table 4.1. Because of this it cannot be meaningful, in this case, to compare the candidates' overall quality through their total marks. The ordering of the totals is simply an artefact of the particular marks chosen to represent the orderings within the papers. In section 7 I shall argue that the ordering of the totals or, perhaps, a nonlinear aggregate of the totals can be meaningful, but only when the examiners have been careful to make and express judgements comparing the qualities of work between the different papers.

The arbitrary choices made in constructing any quantitative model imply that certain mathematical operations (in the example, forming totals) are meaningless when applied within that model: meaningless in the sense that numerical relationships which hold between the results of these operations in one model do not necessarily hold within another equally valid model. We shall say that a mathematical operation is quantitatively meaningful if the numerical relationships which hold between the results of applying the operation hold whatever arbitrary choices are made in the model's construction. The theory of scale types (Stevens, 1951; Pfanzagl, 1968; Roberts, 1979) is a mechanism whereby the quantitative meaningfulness of mathematical operations can be ascertained. Two types of scale will concern us: ordinal and interval. An ordinal scale simply represents an ordering. There are many arbitrary
choices to make in its construction, since any monotonically increasing transformation of it will give an equally valid scale. For an interval scale there are only two arbitrary choices: there is freedom to choose a point on the scale and the unit of measurement. Positive affine transformations are the only ones to give equally valid scales.

For a numerical relationship to be interpretable in qualitative terms, it certainly needs to be quantitatively meaningful. However, a further condition needs to hold. It needs to be semantically meaningful or, as we shall term it, substantively meaningful. The relationship must be interpretable in qualitative terms within the user's perception.

Suppose that the marks assigned on a particular paper lie on an interval scale. Among other things, this implies that mark differences are quantitatively meaningful: the ordering of any two mark differences would be the same in any equally valid model. Thus it is tempting to compare mark differences and say, e.g., that «candidate W's work better than X's by more than Y's better than Z's» simply because their four marks are, respectively, 62, 57, 60 and 58. However, we are not justified in this unless the marks were assigned in such a way that they corresponded to examiners' judgements of this form. Indeed, is not entirely clear that examiners can interpret or make such judgements (French, 1986; French and Vassiloglou, 1986). Again, we defer discussion of this point until Section 7.

For the remainder of this section we indicate the relevance of the above discussion to statistical practice. We shall do so in the context of an example that arises frequently in examination assessment. The issue that will concern us is the appropriateness of parametric statistical inference in certain circumstances.

The papers of Gaito (1980) and Ashby and Townsend (1984) have re-opened the debate on the relevance of the theory of scale types to the practice of statistical inference. The former suggests forcibly that there is no relevance; the second suggests equally forcibly that there is.

Any statistical analysis involves the calculation and usually the comparison of numerical summaries of data. The data might be measured on any type of scale: ordinal, interval, or several others that we have not mentioned. Thus the question of the quantitative meaningfulness of the calculation and comparison of such summaries arises. But is it relevant? Ashby and Townsend, taking up a line of argument begun by Stevens (1951) and Siegel (1956), conclude that it is. If data are measured on an
ordinal scale, means, variances, etc., are quantitatively meaningless and hence any statistical procedure based upon them is inappropriate. Only nonparametric procedures based upon rank orders are appropriate. A necessary condition for the appropriate use of parametric procedures is that the data are measured on, at least, an interval scale.

The opposite view taken by Gaito (1980) builds upon the famous remark by Lord (1953) that «the numbers do not know where they came from». Essentially, procedures of statistical inference and, in particular, hypothesis tests tell us whether there is any evidence that one set of data is drawn from a different distribution from another. Such procedures are based solely upon the distributions of numbers within the two data sets. Whether the data are measured on an ordinal, interval or other type of scale is irrelevant; what matters is the distributions within the data.

To my mind both views are too simplistic in that they seek to make all-embracing statements. They do not recognise that the particular question that an analysis seeks to answer might be relevant to the appropriateness of that analysis. Adams et al (1965) have discussed this point in great generality. Here I simply present two examples based upon ordinal data taken from French et al (1986a). In the first it is clearly inappropriate to use parametric statistics; in the second it might well be appropriate to do so.

Suppose that we consider the marking of a particular essay question. It is conceivable that the marking scheme would simply give a verbal description of the qualities that a candidate must exhibit for a particular mark to be given. For instance, the scheme might take the form shown in Table 4.3.

The important point to note about this type of marking scheme is that it does not necessarily lead to an interval scale of marks. There is neither an attempt to define some concept of «strength of performance», which would lead to an interval scale through difference measurement (Roberts, 1979; French and Vassiloglou, 1986), nor any attempt to add marks, which would lead to an interval scale through conjoint measurement (Roberts, 1979; French, 1981).

Suppose that an examiner uses this marking scheme to mark candidates from two different schools. A question that might be posed is whether there is any evidence that the candidates from one school tended to perform better than the candidates from the other. Because the marks are on an ordinal scale, it is quantitatively and, therefore,
TABLE 4.3

An example of an ordinal marking scheme for an essay

<table>
<thead>
<tr>
<th>Mark</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No serious attempt to answer the question.</td>
</tr>
<tr>
<td>1-5</td>
<td>A few relevant facts, but answer scrappy, no organisation.</td>
</tr>
<tr>
<td>...</td>
<td>..........................................................</td>
</tr>
<tr>
<td>10-12</td>
<td>Solid catalogue of relevant facts; little attempt to muster an argument.</td>
</tr>
<tr>
<td>13-14</td>
<td>A clear attempt at argument, but still a factual essay in the main.</td>
</tr>
<tr>
<td>...</td>
<td>..........................................................</td>
</tr>
<tr>
<td>25</td>
<td>As good an essay as the examiner might be expected to produce in the time allowed.</td>
</tr>
</tbody>
</table>

substantively meaningless to compare average marks. Thus it would be incorrect to translate the question into a hypothesis concerning the difference of two means. Hence, the question cannot be resolved by a parametric hypothesis test. To do so would be to test a worthless hypothesis. However, the question can be interpreted in a substantively meaningful way in terms of whether the candidates from one school tend to be ranked higher than those from the other. Such a question can be investigated by nonparametric rank tests.

Consider, as a second example based upon the same marking scheme, the case when a chief examiner wishes to check that two markers are interpreting the scheme in the same way. Of course, in the final analysis the chief examiner would do this by reading and remarking some of the markers' allocation of scripts; but shortage of time precludes doing this on all but a small sample. So the remarking exercise may be supplemented by a statistical investigation. Suppose that the two markers have been allocated similar populations of candidates, i.e. populations in which the distributions of performances may be expected to be very similar. This judgement of similarity may be supported by evidence from various sources: the candidates' performances on other components, particularly multiple choice tests; teacher estimates predicting the candidates' performances; historical information on the schools concerned; etc. If the markers are interpreting the scheme in the same way, the distributions of their marks should be similar... and this remark holds
whether the mark scale is ordinal, interval or whatever. Thus a natural manner of testing statistically whether there is any evidence of differences in the interpretation of the marking scheme is to test whether the distributions differ. Robustness considerations might well lead one to use a nonparametric test to do this; but, in the event that the distributions can be well modelled by parametric forms, a parametric test could be equally appropriate.

The difference between these two examples should be clear. In the first the question being posed concerns the system that is being measured. Thus any translation of the question into a statistical hypothesis must rely solely on meaningful numerical relations. In the second the question compares two mechanisms for measuring the same system. It simply asks whether two distributions are the same. The first example is very much in the mould of those of Ashby and Townsend (1984); the second is very similar to Lord's (1953) example, much used by Gaito (1980).

5. THE USE OF EXPLORATORY DATA ANALYSIS

In Section 3 I remarked that many of the procedures used in the assessment process are examples of descriptive measurement. Before chief examiners can confidently judge the quality of a candidate's work, they need to understand the context in which it produced and they also need to be assured that the teams of examiners who marked the work were following the marking scheme in the way it was intended. Undoubtedly much of their confidence and understanding develops from their involvement in the design of the question papers and marking schemes, double-marking of samples of the scripts marked by their teams of examiners, etc.; but some confidence can and does develop from statistical analyses of the data. What form should these analyses take? Three remarks seem pertinent.

First, the issues raised and discussed in the previous section arise continually in analysing any examination data. It is safe to assert that marks lie on an ordinal scale, but it is often unclear whether one can assert more; e.g. that they lie on an interval scale. It follows that great care is needed to ensure that analyses are meaningful; they should investigate meaningful questions in meaningful ways.

Second, the problem facing the chief examiners ususally is to recognise trends within the data and, hence, identify aspects that are atypical.
They need to identify questions that were, in some sense, unexpectedly
difficult or easy, candidates who produced work with unusual variations
of quality, examiners who marked in an aberrant fashion, and so on.
Analyses which rely on fitting complicated statistical models may help
in this, providing that considerable emphasis is laid upon residual check-
ing (and such emphasis, in my experience, has been singularly lacking
in practice). However, since the main task before the examiners is to
inspect the data, recognise its main features and identify any relevant
heterogeneity, the potential of simple exploratory data analytic techni-
ques (EDA) should not be ignored. Apart from being designed for the
exploratory investigations, these techniques are usually nonparametric
in spirit. Thus problems associated with the meaningfulness of analyses
are minimised. There is also a third, probably overriding reason for
using EDA.

Examiners are appointed for their expertise in their subject areas
and their general skill in examining. Seldom do they have particular
or, indeed, any expertise in statistical methodology. They are unlikely
to have any feel for import of a correlation coefficient of 0.732 as
opposed to 0.642, say. Undoubtedly, research staff of examination boards
do understand complex statistical analyses summarising the data in
various numeric forms and models, but these analyses will not inform the
majority of examiners. They need clear, easily understood analyses,
which do not require them to be statistically aware to more than a mod-
est degree. EDA techniques with their emphasis on clear, informative,
graphical display of data in ways that facilitate visual comparison
are ideally suited to this need.

These reasons have led myself and others to stress the need for
purpose-built EDA techniques and software for examiners. The British
Secondary Examination Council have provided funds for development
and evaluation of some of these. The remainder of this section describes
a trial in which the power of very simple ways of visually comparing
histograms was apparent. I emphasise that the trial relates to real data
used in assessing a real examination. For reasons of confidentiality,
however, the description does not name the examination concerned nor
are the actual data displayed.

The examination concerned was a humanities subject. Three papers
were set, of which each candidate chose to sit two. Since candidate
choice was involved, it was important that the three papers were of
equal difficulty and marked to equivalent standards». Each paper was divided into three sections. The first section consisted of questions requiring one word answers; the second consisted of short structured questions requiring short sentence answers; and the third consisted of more open ended questions requiring answers of a paragraph or more. The assessment aims indicated that candidates’ work was to be rewarded for factual accuracy, interpretation and critical appraisal. It is clear from the structure of the papers that the marking scheme could only reward interpretation and critical appraisal in the second and predominantly the third section. Marks in the first section could only be awarded for factual accuracy. Each paper was marked by a single examiner, each of whom had drawn up a marking scheme for his paper according to common guidelines as to the total numbers of marks to be given for each of factual accuracy interpretation and critical appraisal.

Figure 5.1 shows the distributions of marks awarded by these examiners in each of the three sections of their papers. Thus the histogram in the top left hand corner shows the number of candidates obtaining each possible mark on section 1 of paper 1. One feature is immediately striking. On papers 1 and 2 the shapes of the histograms are similar in each of the sections. On paper 3 this is not so. There is a trend to a preponderance of low marks in passing from section 1 to section 3. Given the structure of the paper this suggests that interpretation and critical appraisal are not getting there just reward on paper 3. Of course, there is only a suggestion that this is so, and one may counter, as the three examiners immediately did, that the candidates taking paper 3 could well have been weaker than those taking papers 1 and 2. However, Figure 5.2 shows the distributions of marks on papers 1 and 2 conditional on which other paper the candidates took. It can be seen that the population of candidates taking paper 3 performed on paper 1 very similarly to the population taking paper 2. A parallel remark applies to the candidate’s performances on paper 2.

On seeing this evidence, the examiners decided that there was clearly something that needed investigating. They read scripts, discussed the marking schemes and ultimately agreed that the paper 3 examiner had been using rather higher standards to judge interpretation and critical appraisal. Accordingly, action was taken to protect paper 3 candidates from the effects of this in the final award of grades.

Of course, the examiners could have discovered this inconsistency
Figure 5.1. The mark distributions for the sections of each paper.  
N.B. These histograms are only schematic, but they do represent features that were present in the actual data.

of judgemental standards by many other methods. In past years the data had always been inspected by means of numeric summaries. In comparison the examiners said that they found the graphical displays much easier to comprehend; they were able to gain a far better «feel» for the data. Moreover, the change in shape of the distributions on paper 3 had been
by far the most informative aspect. Such change in shape would also be flagged by changes in skewness and kurtosis, as well as effects on the mean and standard deviations; and anyone skilled in statistical analysis would have noted the effect quickly from these. However, the examiners, with due respect, had little more than rudimentary statistical knowledge. For them graphical exploratory analysis had undoubted advantages in conveying
information. Moreover, we should note that it could be protecting them from meaningless comparisons and the need for careful residual checking to identify dubious modelling assumptions.

6. EXCHANGEABILITY AND LATENT TRAIT MODELS

Multiple choice components are quite common in public examinations. They are often called «objective tests», but it will be clear from my Bayesian leanings why I do not adopt this misleading term. In these there are a number of questions or items, for each of which several alternative answers are offered. Candidates must select the answer that they believe to be correct. Usually there is only one correct answer among those listed: in some variants, however, some incorrect answers are designed to be more «sensible» that others and rewarded accordingly: in yet other variants candidates have to select several answers all possessing the same property. Whatever the case, candidates’ scripts take very simple forms, essentially sequences of ticks and crosses. Thus it is not surprising that many statistical models have been developed to describe, analyse and summarise such scripts: see e.g. Lord and Novick (1968) and Weiss and Davison (1981). These models invariably contain candidate parameters which are highly correlated with the number of correct answers which a candidate is expected to give. These parameters have naturally been dubbed abilities; and there is a strong temptation for examiners —or, at least, those who devise assessment procedures— to fit the models to their data, thus estimating candidates’ abilities, and then to grade candidates according to these. Whether they should do so has been a matter of some controversy (see e.g. Goldstein, 1979). Given my remarks in section 2, it will be clear that I side with those who are uncomfortable with such procedures. However, I do believe that such models, interpreted correctly, can be very useful in exploring the data.

Before the examiners see the candidates’ scripts, they have certain expectations. Precisely what expectations will depend on many circumstances, but they include such things as «these questions are of equal difficulty», «performances on these sections should be highly correlated». Since these expectations will influence their judgements, it is important that they are critically examined in the light of the data. To do this the examiners must formulate their expectations as clearly
as possible and they must do so in a language that they understand. Exchangeability conditions can give them a way of doing this.

The data will essentially be a matrix (two-way layout) of responses with the rows corresponding to candidates and the columns to items. Since the examiners usually know nothing about individual candidates a priori, they may hold that any particular data matrix and any row permutation of that matrix to be equally likely: i.e. they may hold the rows to be exchangeable. Occasionally, examiners do have prior information on candidates in the form of teacher predictions of candidates' final grades. In such cases the examiners would presumably only hold the rows to be exchangeable within groups of candidates who have been predicted the same grade by the same teacher. Equally they might also hold the columns to be exchangeable because the items were designed to be of «equal difficulty». In practice, however, unrestricted column exchangeability is unlikely to be reasonable, since examiners usually design papers with a few easy questions at the beginning so that candidates are not disheartened early on. Furthermore, questions are commonly grouped according to subject area. Thus the columns are only likely to be exchangeable within certain groups. Whatever the case, it should be clear that the examiners' expectations can be expressed by exchangeability conditions. Note also that such conditions are easy to explain in everyday language since they involve little more than descriptions of permutations.

Exchangeability conditions have important implications for the form that a person's subjective probability condition can take. The classic result is De Finetti's Representation Theorem (De Finetti, 1937). More recent results are summarised in Aldous (1985) and Diaconis and Freedman (1982). Invariably, these results take the same form. Exchangeability and similar symmetry requirements imply that beliefs should be modelled as mixtures of probability models. The parameters in these models have no physical interpretation. They act simply as indices so that the mixture may be taken and exchangeability ensured.

In the case of examination assessment these results emphasise that, to a Bayesian, ability and other parameters are not quantities that represent objective traits within individual candidates. Rather, they are technical devices to represent exchangeability in the examiners' expectations about the entire population. Thus such parameters and the models in which they lie are relevant when examiners are investigating the
reasonableness of their expectations, but they are irrelevant, except as descriptors of context, when they have to judge individual candidates.

Note also that because of the way that are developed from exchangeability conditions the parameters and models are necessarily quantitatively meaningful, assuming that they are used correctly. The models are automatically developed for contexts similar to the second example based on the marking scheme in Table 4.3.

The precise implications of exchangeability conditions for beliefs about two-way layouts have still to be fully investigated, although some fundamental work has been done (e.g. Aldous, 1981). Also conditions leading to mixtures of Rasch models have been identified (Lauritzen, 1982). Thus the comments and interpretations above should be set against this partial knowledge. None the less, I would claim that their import is clear. Examiners should be asked about their expectations in terms of exchangeability conditions. We have already noted that this may be done in everyday language without recourse to statistical terminology. The implications of their expectations are that certain forms of statistical model should fit the data. This can be tried. If the fit is good, their expectations are reasonable. Any outliers should be investigated, since they indicate either candidates who have performed atypically, and so will need careful, individual consideration, or items which are atypical, perhaps being easily misunderstood by certain categories of candidates. If the fit is poor, the examiners must reconsider their expectations carefully. Exploratory analysis of the data is called for. Throughout it should be possible to communicate with the examiners in easily understood, qualitative terms because the Bayesian framework makes clear the precise, qualitative foundations of the models.

This last point cannot be overemphasised. The examination world abounds with models developed within the classical framework. Unidimensional, numerical parameters supposedly representing objective traits within candidates are found everywhere. Apart from requiring several leaps of philosophical imagination, these models can only be communicated to examiners who have considerable quantitative skills. The Bayesian view, however, builds naturally upon the points made in the previous section. It should be possible to combine the EDA techniques referred to there with more model based analyses to help and guide examiners an the come to understand their data. The software
package currently being developed by Goldstein promises much in this regard (Goldstein, 1986, 1987a, 1987b).

7. MULTI-ATTRIBUTIVE VALUE FUNCTIONS AND MARK AGGREGATION

The previous sections have concentrated on aspect of the role of statistical analysis in examination assessment. Its purpose may not be purely descriptive, but it is predominantly so. The intention is to help the chief examiners understand and monitor the performance of the examination papers, the population of candidates and the teams of examiners involved in the marking. Only when they are confident that they are aware of the important features of these, should they begin judging the performances of individual candidates in order to award grades. Such a time must come, however, and I now wish to turn the discussion to the normative techniques which can support the chief examiners in these judgements.

In French et al (1986a, b) we discuss the forms of these judgements in detail, suggesting that they limen referenced, rather than criterion or norm referenced, as had been argued previously. I shall not repeat that discussion here. I shall only report its conclusion. We believe that the fairest and most consistent way in which chief examiners can make these judgements is, first, to place the candidates’ performances into a rank order of increasing overall merit and, then, identify within this rank order the ranges which correspond to the performances of typical grade «A» candidates, typical grade «B» candidates, etc. In this section I wish to discuss the procedures that the examiners may use to construct the rank order of increasing overall merit.

Examinations typically consist of a hierarchy of components, such as that illustrated in Figure 7.1. The examination is divided into papers; the papers into sections; the sections into questions; and so on. Marking is undertaken at the lowest level of the hierarchy. Marks are awarded to parts of questions, not to complete papers. At the end of the marking process a candidate’s performance is summarised by a vector or profile of marks awarded at the lowest level of the hierarchy. The chief examiners’ task is to consider these mark profiles and place the candidates into a rank order of increasing overall merit. In an ideal world perhaps, they would simply do this by considering each
candidate's profile in turn and slotting it into the rank order entirely by holistic judgement. Indeed, they might not work through the intermediaries of marks: they might simply judge the entire scripts holistically. I say «perhaps» because studies of holistic judgement show it to be fraught with inconsistencies: it is far from «ideal» (Fischer, 1979; Hogarth, 1980; Slovic and Lichtenstein, 1971). However, ideal or not, it does not matter: the sheer volume of candidates precludes the use of holistic judgement in this way.

Current practice «solves» this problem by simply adding up the marks, perhaps with some weighting to allow for the «importance» of certain components. Thus each candidate is assigned a total mark for the examination and placed into the overall rank order according to this total mark. But any function which combines each candidate's profile into a single number will achieve this. Why is it appropriate simply to add? Which in turn begs the question of how do we define «appropriate»?

The chief examiners wish to construct a rank order according to their judgement of the overall merit of the candidates' work. But their judgements are typically not well formed. They may feel that they are prepared to trade off a candidate's poor performance on one paper for a good performance on another, say; but by how much is a question that they might wish to think about.

Our discussion of normative analyses in Section 3 is relevant here. A normative analysis is based upon certain consistency properties or behavioural axioms that an individual accepts as canons of rationality. The analysis compares his judgements with these and identifies any inconsistencies. Any that are found are referred back to him so that
he may modify his judgements. During the process a model is built up representing his judgements subject to the consistency that he is seeking. In the case at hand, the model is a mark aggregation function that forms a single overall mark for each possible profile such that these overall marks rank the candidates as the examiners wish. In constructing this function, the examiners have to consider, among other things, trade offs between performances on the various components. But that is jumping a little far ahead.

First, we need to discuss what behavioural axioms are appropriate. Two are obvious. The overall ranking should be comparable and transitive: it should be possible to position any candidate in the final ranking; and, in pairwise comparison, if the examiners judge one candidate to perform overall better than a second, and the second to perform overall better than a third, then they must judge the first better than the third as well. (French, 1981, 1986.) The third axiom is not so obvious, although it is simple. It belongs to a class of axioms known as *independence conditions* within additive conjoint theory, which lies at the base of all that follows.

Consider the performances of the four candidates shown in Table 7.2. Their marks on four components are shown. These components could be papers or sections or questions. It does not matter; the principles will just the same. For concreteness, I shall assume they are papers. Note that it is generally easier to think of aggregating one level up the hierarchy at a time, even though conceptually it is possible to aggregate directly from question marks to the overall total.

**TABLE 7.2**

An example to illustrate the independence condition

<table>
<thead>
<tr>
<th>Paper</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate A</td>
<td>47</td>
<td>63</td>
<td>45</td>
<td>58</td>
</tr>
<tr>
<td>Candidate B</td>
<td>56</td>
<td>55</td>
<td>45</td>
<td>58</td>
</tr>
<tr>
<td>Candidate C</td>
<td>47</td>
<td>63</td>
<td>73</td>
<td>67</td>
</tr>
<tr>
<td>Candidate D</td>
<td>56</td>
<td>55</td>
<td>73</td>
<td>67</td>
</tr>
</tbody>
</table>
Candidates A and B have identical marks on papers III and IV; their performances differ only on papers I and II. Similarly, C and D have identical marks on papers III and IV and differ only in their performance on papers I and II. Moreover, A and B differ in their performance on I and II in precisely the same manner as C and D. Given this, the examiners might decide that to be consistent they should judge A to have performed better than B overall if and only if they judge C to have performed better than D. In general, they might agree that, if two candidates produce an identical set of marks on a subset of papers, then the overall ranking of these candidates should depend only on their marks on the papers on which they differ and be independent of their marks on the papers on which their performance is the same.

In French (1981, 1985) I indicate some reasons why examiners might reject independence conditions as inappropriate and suggest other weaker conditions that they might adopt, pointing to the theoretical development that would follow. In practice, certain mark aggregation rules are used which do break independence, although it is a moot point whether their users are aware of this. For instance, any rule which selects the best \( r \) out of \( n \) pieces of a candidate's work is suspect. Consider ranking the four candidates in Table 7.2 by summing their best three marks: the result breaks independence.

However, in the majority of examinations independence seems a necessary requirement of the examiners' judgements. Some see it as demanded by the laws of Natural Justice. The quality of two candidates' work should only be differentiated on the basis of aspects that differ.

If independence is accepted, then a very simple result follows. The mark aggregation that should be used has the form:

\[
\sum_{i=1}^{n} v_i(x_i),
\]

where \( n \) is the number of components to be aggregated and \( x_i \) is the mark on the \( i \)th component. The functions \( v_i(\cdot) \) essentially bring the \( n \) components' marks onto a common scale. Strictly, the result only follows if certain other «housekeeping» axioms are accepted. These are noncontentious in the examining context. They are discussed and described in French (1981, 1986) and Roberts (1979), along with multi-attribute value theory and the more general additive conjoint theory from which this result derives.
The \( v(f) \) are assessed by asking the examiners how much a mark on one component is «worth» in terms of marks on others. Thus the examiners are asked to consider the tradeoffs of marks between components. These tradeoffs need not be constant: the tradeoff of marks between two components may depend on the general levels of marks on those components. Thus the \( v(f) \) may be nonlinear functions. For instance, in a science examination, practical and theory marks may be equivalent for discriminating poor performances, but theory marks may be deemed more important in discriminating between better quality work.

Details of possible assessment procedures of the \( v(f) \) are given in many places in the literature. French and Vassiloglou (1986) describe the «lockstep» procedure in the context of aggregating two paper marks. Keeney and Raiffa (1976) and Fishburn (1967) describe other procedures, albeit in the context of multi-attribute value assessment.

Given my remarks in Section 4 I should enlarge upon the substantive meaning of tradeoffs between components. What qualitative judgments underlie them? The examiners are asked questions of the following form.

«Consider two candidates with mark profiles \((\_, \_, 47, \_, 56, \_)\) and \((\_, \_, 50, \_, X, \_)\). Both candidates share the same marks on components other than the two indicated. The second candidate has a mark of 50 on a component for which the first has 47. What value of \( X \) would lead you to judge that the two profiles represent equivalent performances? Presumably \( X < 56 \), but by how much?».

[Note how the assumption of independence is used in framing this question. The common performance on the other four components is irrelevant.]

Questions such as these require the examiners to compare the relative strengths of performances represented by marks on different components. It is shown in Vassiloglou and French (1982) and French and Vassiloglou (1986) that only by making these qualitative judgements can the examiners be sure that the components are aggregated in a meaningful way. In the first of the papers we argue from Arrow’s Theorem that such tradeoff information is necessary in order to be fair and consistent to the candidates. It is not sufficient simply to aggregate rank order information from the components. As the first example from Section 4 suggested, there must an attempt to compare marks across com-
ponents before an aggregate can be formed meaningfully. This holds even if the marks within components correspond to equal intervals in some strength of performance notion — i.e. the difference in performance that the examiners perceive between, say, work awarded 81 marks and work awarded 82 is precisely the same difference as that between, say, 36 and 35. Aggregation rules must be based upon judgemental comparison across components.

It should be noted that, while the assessment process ensures that the \( f_A(\cdot) \) bring the marks on the various components onto a common scale, there is no suggestion that this is an absolute scale. The aggregated marks only have meaning in the context of the particular examination. To compare marks between different examinations the examiners must make further trade off judgements between these. Such judgements are required of the examiners when they are asked to ensure that the grades awarded are equivalent in standard to those awarded in previous years.

Earlier in the paper I argued that normative analyses bring understanding to the user by making him reflect upon his judgements and so guiding their evolution. The questioning of an examiner about his trade-offs does precisely this. It helps him reflect upon the relative worth of marks earned in different parts of a candidate’s work. Moreover, since the trade-offs between one component and a second and between the second and a third imply the trade-offs between the first and the third, it is possible to investigate the consistency of his judgements and invite him to reflect upon and revise any inconsistencies.

I began the section by remarking that current practice simply adds marks. This may be an appropriate thing to do, but only if all the assessment procedures have determined that the \( f_A(\cdot) \) should be linear. In well planned examinations with marking schemes designed with some care to ensure that each mark rewards the same unit of performance on every component, this may be so. However, there is no guarantee: it must be checked in each case.

Many other aspects of mark aggregation functions are discussed in the literature. I would refer you particularly to Vassiloglou (1984), since that paper not only translates multi-attribute value/additive conjoint theory to the context of examinations, it also develops new theory.

Is the theory being applied? In the strict sense, no. Part of the research project referred to earlier is the develop the theory into practical
software for evaluation. This currently underway. However, these ideas are having practical effects. They are focussing debate and raising issues that have long been ignored. The judgements required of examiners have seldom been recognised explicitly. Now they are, as is the need to find tools to support examiners in their judgements.

ACKNOWLEDGEMENTS

The ideas in this paper developed in joint work with John Slater, Marilena Vassiloglou and Alan Willmott and in many places the paper draws upon work in our joint publications, especially French et al (1986a, b) and French and Vassiloglou (1986). Our work has been funded by the British Secondary Examination Council. An early version of the paper was presented at the annual meeting of the Spanish Statistical and Operational Research Society in Gijon in September 1985. An early version of section 6 was presented in a contributed paper at the conference held in memory of De Finetti in Innsbruck, September 1986 (French, 1987).

REFERENCES


REPLY TO DISCUSSANTS

I am grateful to my discussants for the comments: clearly, they do not need exhorting to «think about things». I must confess, however, that in some cases their thoughts have led them to somewhat different conclusions from my own.

I guess that in many respects the point of departure for many of our different perspectives is the purpose that we ascribe to public examinations, as represented by the English and Welsh GCE A level system. I do not believe that their purpose is to measure the candidates’ abilities, achievements, or whatever — at least not in the sense generally accorded to the activity of measurement.
Before I argue this further let me make one point. I do accept that most of the public and the majority of examiners would, if asked, differ with me on this. But I do not accept that their responses would be other than «knee-jerk» replies conditioned by a comforting fiction which has evolved over the past century or so as our examining system itself has evolved. Furthermore, I accept, as Dr. Forrest suggests, that some of the procedures adopted by our examination boards do not sit comfortably in the framework that I propose. But, while I have much respect for what the boards do, there are a few procedures that I would change.

To me the claim that one can measure the ability, intelligence or whatever of a person is to suggest that one believes that with some suitable «scalpel» one can dig these entities out of the brain or wherever. Psychometric tests are often supposed to be such a «scalpel»: but just as medieval anatomists failed in their quests, so I expect the psychometricians to fail. Where is my evidence to support this assertion? I freely admit that I have none, save my total inability to construct a self-consistent view of educational assessment (and psychometrics) based on the objective existence of some measurable intellectual qualities of individuals. Every time I try, I create, for me, incredible fictions. Nor, when I read the literature, do I find that others have been more successful. Unlike Dr. Biggins and Professor Loynes, I cannot in conscience accept pragmatically what my logic emphatically denies. What I can do is rationalise and construct a system of assessment based upon the foundation that it is the judgements of examiners which are reported. That I, and others, have tried to do here and elsewhere (French et al., 1986a, b).

Whereas I believe that physical objects exist and have lengths, etc. which can be measured, I believe that intellectual qualities such as intelligence are mental constructs made by observers to explain the behaviour of others. Over time, a general consensus over them has arisen so that judgements of these common constructs are generally «in line». But not always.

There are many tales, not apocryphal but true, of candidates who, within a matter of days, take examinations in the same subject set by two examination boards and obtain quite different grades, in some cases failing outright at one board but passing comfortably at the other. The common explanation for this is something akin to «measurement error». The inherent day to day variability in the candidate’s skill in responding
to examination questions combined with the «marking errors» of the examiners, kept to a minimum by the boards’ procedures but still present, are supposed to have led to the difference in the candidate’s results. I am sure that these two causes of the difference in grades are present. But I do not believe that they are the only cause, nor perhaps the dominant one.

There is another explanation. I have argued that there is no single entity within the candidate that can be measured. There are only examiners’ judgements of the «quality» of the candidate’s work. The different examiners at the different boards differ, perhaps, in the attributes that they judge to determine quality or in the weight that they ascribe to the attributes. In short, the consensus between the examiners at the two boards falls far short of perfect. Thus I would turn Prof. Jackson’s point back on him. I can use my subjectivism and belief in a general, but not perfect consensus to «explain why groups of persons separated in time and/or space» do not always «share the same perceptions».

I am in danger of spending too long on this issue and not addressing the other points my discussants raise.

Professor Townsend suggests that I may still be missing some subtleties in the relevance of scale types. Perhaps so, although I do believe that questions concerning differences in mechanisms by which numbers are attached to objects are quite distinct from questions that concern differences between the objects that may be represented by differences in the attached numbers. Moreover, the statistical analyses appropriate to one case may — and probably do — differ from those appropriate to the other. However, to dwell on that discussion is to deny emphasis to the agreement between us. One should think carefully about the scale type of one’s data before analysing it: much more carefully than many of us seem to do. Despite the seeming stronger disagreement with Professor Loynes and Dr. Jackson, I think they too are arguing this and our problem, here, is one of communication.

Dr. Biggins and Prof. Loynes both observe that marks may well be assessed on more than an ordinal scale: I agree. But the point that I and Professor Townsend would make is that this should be examined in each and every case. I wish I was as confident as Dr. Jackson in his suggestion that «any respectable introductory course on the statistics of educational testing» makes this and many other of my points. I suppose
that the word *respectable* makes this a truism, but then I would deny respectability to rather a lot of courses and much practice.

Prof. Loynes takes me to task for not providing evidence that the derivation of parametric models from exchangeability assumptions implies that the parameters have no physical interpretation. He misses the point. There is, on the contrary, no evidence that they do. Bayesians are careful to derive their models, where possible, from clearly stated assumptions. Exchangeability ideas are providing us with a mechanism for achieving this. Exchangeability requires that the observer thinks carefully about symmetry in his or her a priori beliefs concerning a sequence of observations. All the assumptions are about the sequence. The models, therefore, describe the observer's view of the sequence. To extract parameters from the model and attach them to individual observables, tempting though that may be, is unjustified.

I am not the most perfect of communicators and I seldom choose my words as precisely as I should. Professor Loynes notes that I often unwisely use words which have connotations that I deny. Alas, yes. More importantly, at times I find myself thinking of the discussion: «but I thought I had said that». There are places where I am less sure of our differences than my discussants seem. This is particularly true when I read Dr. Forrest’s comments. Unlike the rest of us, Dr. Forrest works full time in the world of public examinations. We comment: he does. If I felt that my prescription was leading me far from the procedures that he would use, then I would have serious doubts as to its validity. I am comforted that he believes that examination boards should seriously consider adopting the methods I propose.

Lastly — I cannot resist it! — I totally agree with Prof. Loynes that «any responsible statistician ... thinks carefully about certain things»: but then any responsible statistician would be a Bayesian, wouldn’t he or she?