TRABAJOS DE ESTADISTICA
Vol. 5, núm. 2, 1990, pp. 95 a 117

COMENTARIOS AL ARTICULO DEL PROF. FRENCH (Vol. 4, n.º 1, 1989 de *Trabajos de Estadística*)

En el artículo del Prof. French, titulado: STATISTICAL AND DECISION THEORETIC ASPECTS OF EXAMINATION ASSESMENT (vol. 4, núm. 1, 1989, de *Trabajos de Estadística*) no se publicaron los comentarios que, al citado trabajo, formularon diversos autores.

Al objeto de completar dicho artículo se publican a continuación tales comentarios y se repite la contestación del autor a los mismos.

COMMENTS BY PROF. D. D. BIGGINS

Dpt. of Probability & Statistics,
The University of Sheffield,
Sheffield S10 2TN, U. K.

If we are going to examine then we should endeavour to examine properly, and we cannot then escape spending time thinking about what we are doing. I found Dr. French's article on the matter thought provoking and therefore interesting and valuable. Among the issues it stimulated me to think about were the following.

1. One does not have to believe that planets are point masses to use this concept and find it helpful in modelling planetary motion. Neither does one have to believe that there is a «true ability» which is «inside each person» for it to be a useful modelling concept. (Here I do not intend modelling to be interpreted in the narrow mathematical sense.) In any case I do not find the contrast between those who supposedly believe in «true ability» and those who do not, but are prepared to admit concept(s) like the «overall quality of the scripts of a candidate» or his or her «overall merit», as stark as is maintained in this article. For modelling purposes these latter concepts will often serve just as well as «true ability».

What mi one mean by «true ability» anyway? Certainly the interest in examination results is generally because of their supposed correlation with performance in other tasks, and without such a supposition it is hard to see why anyone would care about the results. I wonder whether talk about «true ability» cannot be viewed as a presumption that such correlations do (or perhaps, should) indeed exist.

2. The exhortation to use graphical methods which are comprehensible to the examiners (which is my interpretation of EDA here) is one I wholeheartedly endorse. Such methods may be intented to illustrate the conclusions of more sophisticated numerical procedures, and may need to be supplemented by numerical summaries, but their value in aiding understanding cannot be doubted. In the author's example, where there are many candidates, few papers and few examiners, histograms and, I assume, scatterplots (or some other indication of the correlation structure) provide a reasonable «non-parametric» graphical analysis. However, once there are many papers and fewer candidates (each taking a selection of the papers), I see no real alternative to a more «parametric» procedure. Here by parametric I mean that one will need to be prepared to do more to marks (e.g., calculate averages) to construct the plots. For example, a possible display in this situation is a scatterplot for each paper of the marks against each candidate's average mark on the papers he or she took. Of course, if marks are no more than ordinal this will not be a reasonable thing to do.

3. Are marks on an ordinal scale only? I rather doubt it. If they are then *any* monotonic transformation of the marks will lead to an equally valid mark assignment. I think this condition will rarely be fulfilled, for the marks generally have a context that influences what is reasonable as an assignment, and monotonic transformations that have a slope that is too variable may well lead to unacceptable marks. This seems to apply even to marking schemes of the form given in Table 4.3. I am not trying to maintain that the marks are on an interval scale, but only that they usually do contain more than ordinal information.

4. I think Dr. French's proposed procedure is (having obtained the marks) to (i) perform an appropriate descriptive analysis (what is appropriate will be determined by the context), (ii) reflect on the results of this until the shape of the marks is well understood, (iii) select a

suitable transformation for each set of marks and average (or total) the results to ranks the candidates. If possible this does seem to be a reasonable programme, though there are many reasons that may make step (ii) rather difficult. Whilst the example described in section 5 seems ideal in the opportunity it affords at this stage, it is not uncommon for conclusions to have to be reached rather quickly and by a fairly large group of people, none of whom have the expertise, or time, to have a proper overview.

Also, the context in stage (i) includes expectations about what kind of combination of the marks will measure the overall quality of the scripts. In particular there may well be restrictions on the kind of function v_i (introduced in section 7) that are thought reasonable. Furthermore, an interim judgement about the overall quality of each of the students may be precisely the information that is used in forming an impression of what constitutes various levels of quality on a particular paper, or, in a more extreme case, in picking out a paper as aberrant. Consequently I would wish to allow the relationship between stages (i) and (iii) to be more interactive than the description above suggests. In particular I think it makes sense to use algorithms in the choice of the v_i which select them from within some class of functions deemed reasoable, in a way which tries to reflect expectations about how overall quality will be manifest in the component papers. I would, of course, still insist on some version of stage (iii), as a final step in the process, for, in this area, the results of any algorithm must always be understood properly and considered critically rather than being used automatically.

COMMENTS BY C. M. FORREST

Joint Matriculation Board
Manchester, England

It is very appropriate and timely that Simon French urges the examining boards in England and Wales to check on the validity of their awarding procedures. In recent years there have been many changes to the public examination system but, like Caius Petronius [1], we have to ask ourselves whether or not these changes have materially affected what it is we do. Certainly on the face of it, the examining system is very different than what it was thirty years ago. In the General Certificate of Education (GCE) entries to the Advanced level (originally meant for eighteen year-olds after two years in the sixth form) have nearly quadrupled; the proportion of female candidates has gone from less than one-third to almost one-half; the proportion of adults offering the examination has increased; the proportion of candidates offering Arts subjects has steadily declined, but not as steadily as Science subjects, whereas that for the Social Science subjects has increased enormously. Then there has been the inclusion of a common core in the syllabuses of some major subjects, common that is to each of the boards; and most recently of all there was in 1987 a change in the grading scale used for the reporting of results. At GCE Ordinary level (normally taken at 16 years of age) some of the same sort of trends may be found, but here the situation is even more complicated. In 1963 the Certificate of Secondary Education (CSE) was introduced, also for 16 year olds. From 1988 these two examinations are to be replaced by the General Certificate of Secondary Education (GCSE) and it is estimated

[1] Caius Petronius was a Roman writer who, in the first century AD, complained as follows. «We trained hard, but it seemed that every time we were beginning to form up into teams we would be re-organised. I was to learn, later in life, that we tend to meet any new situation by re-organising, and a wonderful method it can be for creating the illusion of progress while producing confusion, inefficiency and demoralisation.»

that perhaps as many as 650000 candidates will sit the examination for GCSE in 1988 making between them about four million subject entries.

Thirty years ago GCE examinations, all on a single subject basis (which they still are), were relatively simple, the majority of question papers consisting of a choice of questions where the candidate was required to respond in continuous prose (that is, they were «essay questions»). Now many candidates are required in addition to respond to short-answer questions and objectively-marked questions; moreover there may be practical examinations.

Teachers' assessments of coursework, of practical skills, of the «process» rather than of a «product», feature widely and in GCSE will be a requirement. Greater use of oral assessments is not far away, even in subjects where so far it has not been a common practice. The one thing that has not changed, however, is the almost invariable practice[2] of adding the marks from the various components in order to produce an order of merit (in terms of total marks) on which distribution the grades are imposed.

The practice of providing results on the basis of total marks is as old as public examinations themselves, going back over a hundred years. Current attempts to move towards a system of awarding grades in terms of criteria (in which a specific grade would only be given if the candidate has achieved the predetermined criteria for that grade) are, not unexpectedly, having to cope with enormous problems and it would de unrealistic to believe that a fully-fledged awarding process involving criteria can be worked out before the end of the present century. Thus there may well be considerable advantage in giving the ideas put forward by Simon French serious consideration.

His first question is effectively «Why do we add marks?». It is no answer to say «We have always done this» even if it is true. Certainly in the early days of public examinations it seemed perfectly logical to sum a candidate's marks from each of the questions answered and then to use that aggregated mark as the measure to decide whether or not the

[2] Although the practices of the nine GCE boards in England, Wales and Northern Ireland «differ in detail... what is attempted is essentially the same; at each critical borderline, experienced examiners decide where the threshold sould come on the mark distribution in the light of the evidence gained from a study of the candidates' scripts, the question papers, the known characteristics of the entry, background knowledge and statistical information provided by the staffs of the boards» (*Bardell, Forrest and Shoesmith*, 1978, pages 12-13).

candidate could be deemed to have passed or failed, or be selected for the Civil Service or university. It is true however that Edgeworth (1890) made the suggestion that it might be better to combine marks by taking their geometric mean rather than their sum.

As Simon French points out «Psychometric concepts and theories» have been, and are, applied to the whole range of measures made in schools and colleges throughout the world, including attainment examinations like GCE Advanced level. We must not forget however that examinations of this kind were on the go years before the theories came to be formulated. (And it is well to remember that sometimes the educationist is not too particular in keeping to the strict assumptions which underlie the statistical methodologies used!)

Be that as it may, French's argument logically starts from the point where most of us are: «the feeling that the purpose of examinations is to gather data form which some "level of achievement in the subject examined" may be estimated» (pages 7-8). But French rejects this. «I do not believe that there is any entity within a candidate that could be measured in such a way. Instead the purpose is to report the judgements of the examiners. Moreover, the examiners do not make their judgements about some trait that they postulate to exist within the candidate...» He goes on. «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» (page 9).

While it is possible to raise queries about some of the phrases which appear in the above extract, so to do may be counter-productive in that attention would neccessarily be drawn away from the central issue, that is, that marks do not «encode an entity withing a candidate» (page 9).

French tells us that the «purpose of a public examination is not to measure in some objetive sense something directly about a candidate» (page 9). But who says it is? Almost every text on educational measurement, and usually in an introductory section, stresses that measures of an educational (and psychological) kind will be indirect, unlike the situation in the sphere of physical measurement where, for example, a person's height is measured directly by reference to a scale of length. There can be no direct measure, however, of a candidate's attainment in a particular subject. Moreover, Lindquist (1951), lists six obstacles

which prevent the direct measurement of educational achievement (pages 142-144).

If French eschews the idea that marks «encode a measurement of an entity within the candidate» (page 9), he believes that they «encode the examiner's judgement of the quality of the candidate's performance (page 9). This view is difficult to sustain given the way most boards currently organise their Advanced level examinations. Most, if not all, subjects involve more than one component and each of these is marked by a separate panel of examiners. Even in subjects where an examine marks all the work of an individual candidate, that examiner is unlikely to be required to mark all candidates' work. (Perhaps it is worth noting here that examining boards tend to use the word «examiner» in two senses: first, for those senior persons who are concerned with the overall picture, being responsible not only for the construction of the question papers and mark schemes but also for the decisions on major grade borderlines. Second, boards refer to as «examiners» those who, under supervision, are responsible only for marking candidates' answers.) It is difficult to see how French can suggest that any examiner can be in a position to make a judgement about the quality of a candidate's performance.

Yet there is, in French's ideas, sufficient for the examining boards to consider seriously whether or not the time has now come for them to take the radical step of not using a candidate's aggregate mark to determine the grade but to adopt the idea of awarding grades in terms of the patterns of candidate's component marks. So to do would not imply the operation of fixed hurdles on one or more of the components. The computer software produced by French and his colleagues allows the use of what have been called «soft» hurdles: the simple lack of one mark in a particular component does not necessarily condemn a candidate to a low grade —the judgements of the examiners in formulating the rules for combining component marks allows freedom for much subtlety.

Any recommendation to the examining boards that they consider such a change will have to be framed in far more simple and in less technical terms than in the article now being discussed. The boards too must be assured that it will still be possible for them to claim that this year's standard is the same as last year's — and this is where senior examiner judgement is vital. In practical terms it may well be advisable

to suggest that initially only those subjects with large entries be involved so as to allow experience to be built up gradually and confidently.

The introduction of criteria for the awar of grades in GCSE is proving difficult, as might have been foreseen. Nevertheless, movements of this kind at 16+ examinations will have an effect on 18+ examinations. The use of examiner judgement to determine GCE advanced level grades by a consideration of component mark patterns can be seen as an important first step in that direction.

References

BARDELL, G. S.; FORREST, G. M., and SHOESMITH, D. J. (1978): *Comparability in GCE: a review of the Boards' studies, 1964-1977*», Manchester, Joint Matriculation Board on behalf of the GCE Examining Boards.

EDGEWORTH, F. Y. (1890): «The Element of Chance in Competitive Examinations», *Journal of the Royal Statistical Society,* Vol. LIII.

LINDQUIST, E. F. (1951): «Preliminary Considerations in Objective Test Construction», in Lindquist, E. F. (Ed.), *Educational Measurement,* Washington, DC, American Council on Education.

(The views expressed in this article are those of the writer and not necessarily those of the Joint Matriculation Board.)

COMMENTS BY DR. P. JACKSON

Department of Statistics
University College of Wales
Penglais-Aberystwyth, SY23 3BZ, U.K.

This paper, the latest sortie in Simon French's long campaign to wrest control of the domain of public examinations from the Lord-ship of psychometrics and annex it to fiefdom of preference-rating theory, contains a strange mixture of sound common sense and highly controversial philosophical speculation.

That one should be aware that many educational scoring procedures yield only an ordinal scale, that one should examine one's reasons for treating a particular rescaling as though it were an interval scale, that one should study the overall outcome of one's procedure to see whether it conforms to prior expectation, that data should be examined graphically before being made the subject of complex analyses, all these are points made in any respectable introductory course on the statistics of educational testing.

However, to derive from the admitted difficulties of achieving reliable measurement in the social sciences the conclusion that there is nothing there to be measured, so that a GCE A-level examination has the same logical status as a beauty contest, is surely a classic case of throwing out the baby with the bath water. One might as well say that in primitive communities without thermometers, temperature does not exist.

There is, of course, a philosophical precedent for this view: the argument that because all knowledge comes to us through our perceptions, we are entitled to discuss only those perceptions and not to infer a reality «out there». The practical difficulty with this position has always been to explain why groups of persons separated in time and/or space share the same perceptions.

If an examination result is merely an expression of the examiners'

preferences, there is no obvious reason why anyone else should be interested in it. In practice, examination results are believed to have value as predictors of future achievement. Since it is the candidate, not the examiners, who moves from school to university or employment, it seems more logical to say that the examiners have succeeded in measuring (or at least assessing) some characteristic(s) of the candidate, than to say merely that they have achieved a coherent way of expressing their preferences.

The effects of the philosophical presupposition begin to show in Section 4. In discussing the use of parametric inference procedures the author cites two questions which might be asked about scores on a test which yields only an ordinal scale: (i) whether school A is performing better than school B, for which he prescribes a non-parametric test; and (ii) whether examiner A is marking on the same scale as examiner B, for which he allows that a parametric test might be appropriate. The difference he ascribes to the first being a question about the candidates being marked and the second a question about the marking mechanism. But surely the difference lies in the question being posed? If one were to ask (i) whether school A has the same spread of abilities as school B; and (ii) whether examiner A is being more generous than examiner B, the argument would be reversed.

The philosophic presupposition presumably also accounts for the absence of any discussion of measurement error, meaning, in context, the fact that candidates will not score identical number-right marks on groups of items which the examiners, even after careful scrutiny, regard as exchangeable. As a result, Section 7 does not address the question of whether the examiners might use their prior beliefs about relationships among the components to judge where such «errors» have occurred, and to moderate their judgements accordingly.

A brave display on the field, but not a convincing victory.

COMMENTS BY PROF. R. M. LOYNES

Dpt. of Probability & Statistics,
The University of Sheffield,
Sheffield S3 7RH, U. K.

Virtually everyone engaged in education at any level ends up examining students and is therefore, at least indirectly, involved at least in deciding how to treat the results of these examinations. Yet surprisingly enough it seems that the number of those who think in any depth about the process, as opposed to about the details of the marks for individual questions, is quite small. The present paper therefore provides a welcome opportunity to consider some of these broader aspects.

But although I think that a great deal of attention must be paid to these concerns, and in that sense I agree with Dr. French, I find that quite a lot of the detail is something that I cannot agree with. The first point I would like to make is a very general one about the question of the Bayesian viewpoint. I am not myself a Bayesian, though I am not in any real sense anti-Bayesian, and indeed I have been known to tell my students that I would be a Bayesian if only I knew how. But I do not think that some of the claims made for Bayesianism, implicitly at least, are really justifiable. For example in the first paragraph of Section 1 the last sentence explains that a Bayesian is someone who thinks carefully about certain things: I rather think that any responsible statistician of any school of thought would believe the same thing.

However, there are more important matters at issue here. The most significant message contained in the paper is surely that there is no such thing as an underlying ability to be measured by examinations. I suppose this is a logically defensible position but it does seem to me to lead one inevitably to ask what exactly the point of reporting the performance is. I cannot believe that the very large number of parents, of employers, and of others who regularly inspect the grades which are given as a result of public examinations do not themselves believe that

there is some reflection of an ability in this. I am not even entirely convinced that Dr. French himself is completely committed to this view. There are various words which appear in the text which suggest that at the very least he has not been able to remove these ideas from his language even if he has from his mind. For example, there are various references to a fair procedure and this seems to be distinguished from the idea of a consistent procedure, which I take it would mean one which was the same for one candidate as for another. There is also a reference in Section 1 to examining boards seeking to award final subject grades in such a way as to have meaning to those receiving and using them. It seems to me that if all one means by this is that the grade represents the results of the examination then one is really doing no more that uttering a tautology.

I find myself in disagreement with quite a number of details in addition. For example, just after Table 4.1 there is a remark that the first instinct of examiners is to compare the totals. I think that is rather a common instinct but I am not convinced it is quite as universal as is implied there: for example, some examiners would look at such quantities as the median mark in order to get some idea of candidate performance. Then again, just a little later, there is a remark that the quantitative relation between 65 and 63 on paper one represents the ordinal judgement that B's work was better. I am not convinced that all examiners would agree this was an ordinal judgement only. It is very difficult to make precise what is meant by a particular mark but most examiners would, I believe, regard the comparison between 63 and 65 as suggesting that there is no very great difference in quality, whereas if it was a question of comparing 63 with 93 there is a substancial difference; this certainly does not amount to believing that marks are on an interval scale but it does seem to involve rather more than an ordinal one, and I think that this is one of the difficulties that arise when one talks about types of scale. But I also think that the arguments about scale have been taken further than they warrant. For example, in Section 4 the distinction is drawn between testing whether candidates from one school tend to perform better than those from another, and comparing two sets of marks to assess examiners. The position is that in the first case it is not appropriate to use averages whereas it is quite reasonable in the second, but I fail to follow the argument. Clearly any attempt to use averages to estimate must be invalid, as must any

attempt to test for a (specific) non-zero difference in the mean: the averages, or what they measure, are meaningless. But if it is a question, as it seems to be here, of testing for no difference between populations, which will certainly imply that averages are equal excepts for statistical variation, then it seems to me to be clearly valid to use averages. (Considerations of power, or difficulties in dealing with the distribution of the test statistic on the null hypothesis, may lead one to choose different test statistics, however.) Incidentally, I presume that in this context, where marks mean nothing more than a representation of the actual performance, there can be no room for measurement errors or errors of a similar kind; this does make me feel uncomfortable I am afraid. There is a reference in Section 5 to marks which says that it is safe to assert that marks lie on an ordinal scale. Partly because of this problem of errors, which I feel I need to retain, I would much prefer to talk about lying on an approximately ordinal scale.

To take up a different point, it seems to me that there are one or two places in the text where the argument is incomplete: where the conclusion just does not follow from the premises. In Section 6, for example, about exchangeability there is a statement which says that the parameters in these models have no physical interpretation. No justification is given for this, but then a line or two later the theme is taken up again with the remark that the results emphasize that, to a Bayesian, ability and other parameters are not quantities that represent objective traits within individual candidates. I do not see what the evidence is for drawing these conclusions. At best the argument could be that such parameters *need* not represent real quantities, not that they *cannot* do so, at which point one starts to become interested in questions of what is reality. Similarly further on in the same section there is a statement that the Bayesian framework makes clear the precise qualitative foundations for the models. It seems to me that the wording would be more defensible if it said that it made clear some possible precise foundations.

Moving on to section 7, and the axioms that one might wish to adopt, I am not myself entirely convinced that the ranking should be comparable, or rather that every pair of candidates should be comparable. Certainly in practice one could very well imagine a situation in which one was neither willing to say that two candidates' performances differ, nor that they are the same. Perhaps this is only to do with ordinary human uncertainty in approaching an ideal, but perhaps also it should

be taken into account; I remain uncertain of this. Then the axiom of independence: at first sight this does seem fairly convincing but suppose one imagines the following example:

	Analysis 1	French	Analysis 2
A	30	51	70
B	32	50	70
C	30	51	10
D	32	50	10

It seems to me that here, just because the first and third papers depend on much the same kind of ability, one might argue that in comparing A and B one can see that both are good at analysis and that one need not worry so much about the first paper, and therefore go on to conclude that A is better overall than B, since it is then the second (and third) paper that we are going to care about; on the other hand, when we come to candidates C and D it is not at all clear by looking at the third paper that they know any Mathematics at all, in which case one might, it seems to me, perfectly consistently argue that in this case D is actually better than C. The axiom of independence perhaps depends therefore on the papers being independent, or at least equally dependent, which introduces an awkward problem of judgement into the discussion. For a different kind of example one has only to imagine a practical test of competence in which one demands a minimum level in each one of a number of subjects. This clearly does not satisfy the axiom of independence and yet in the right circumstances it is obviously an acceptable thing to expect.

To come back to the general point, it seems to me that there are good reasons for considering that there is a real ability underlying performance in the examination, although it may very well be a complicated and compound ability as was suggested in the quotation that appears in Section 1. In particular without some such assumption there seems to be no justification for looking for unusual individual cases, i.e. outliers, nor indeed in the end for dismissing any proposed method of combining marks as unreasonable.

COMMENTS BY JAMES T. TOWNSEND

Purdue University
W. Lafayette IN47907, USA

The worthy Professor French has given us much food for thought in this pellucid monograph on procedures in assessment. For the most part, there is little with which I could find cause to take serious issue. But, there are two main areas where he and I apparently disagree. The first is on the exact interpretation and use of types of measurement scales. This one is, I believe, relatively minor because unlike, say, Gaito (1980), Dr. French seems to at least believe in the distinction among scales and that some care should be taken in the application of statistical techniques. The second is perhaps more interesting in a scientific sense; it is certainly more general. It concerns the past, current and future epistemology and rationale for assessment through psychometric test procedures.

Is scale type ever irrelevant?

Though in some sense a «nicety», the argument is, I believe, of some import. Let us take up the example to which Dr. French refers as an example of a case where the underlying scale is presumably of no consequence. Two markers are given materials from two more or less identical populations. For the sake of argument, let us suppose they are indeed identical. The question is «do the markers grade the tests in the same fashion»? This translates into, «do the distributions of the two markers differ»?

Now, one might cut the Gordian Knot immediately by assuming that the underlying scales were but nominal. Then the very question is nonsensical, unless the investigator notices the nominality and interprets it in an appropriate sense, as with apposite multinomial assays. One must then perforce conclude that scale type is in fact important.

But, let us look for more subtle disputation, by assuming that the

scales are ordinal whereas the investigator employs techniques that rely on interval strength. An example would be to test between the means using a t-or z-test. The critical question is whether a critical error could be made because of the scale confusion. The answer is «yes». Suppose the distributions are not identical, yet the means are not statistically different at a usual level of significance. The investigators concludes that the distributions do not differ. The fact that the means are not different may easily be changed by ordinal transformation (see, e.g., Townsend and Ashby, 1984), in which case one might conclude that there was a difference, but from a totally erroneous bassis.

That takes care of Type II errors. What about Type I errors, that is, where the investigator mistakenly concludes that a difference is present? At first glance, this would seem impossible: If the distributions are the same then that is all there is to it, and no mistake could be made using any relevant statistic! However, it is not really so simple, because of the presence of noise in any sample data (i.e., the sample of the markers' assessment). The presence of a difference in the means is meaningless (no pun intended) because a perfectly legitimate transformation gets rid of it, and we still do not know whether the markers are behaving in a sufficiently similar fashion.

No, I am afraid that one must perform tests in the above circumstances that are valid within the ordinal framework. A good method for the present hypothetical case might be the Kolmogorov-Smirnov test of stochastic ordering (see, e.g., also Townsend and Ashby, 1978; Townsend and Ashby, 1983, Chapter 8; for advantages of such methods even in the presence of ratio scales).
This ordinal test performs just the service called for, and in the bargain, if the distributions are ordered, then it follows inexorably that the means themselves are also ordered.

As a final comment on a somewhat related topic, I would like to suggest, following the excellent advice of Maurice G. Kendall and Alan Stuart (1973), that we take care to discern the difference between «nonparametric tests» (i.e., parameter free tests) and «distribution free» tests. The former does not necessarily hold outside a particular family of distributions, for instance, the normal family. The latter should be good either throughout the universe of distributions, or at least a broadly specified class of families, for example, «all absolutely continuous distributions, unbounded on the right».

What is, or should be, the ultimate purpose and methodology of psychological testing (and in particular public examinations)?

In my opinion, it is entirely fair for the social scientist, or statistician for that matter, to focus either on the examiners (markers and so on) or alternatively on what may be going on «in the heads» of the testees. As a psychologist, I must profess an interest in the internal workings of the «psyche», to risk derision with a rather out-of-fashion term.

From this point of view, a major problem, indeed an abstruction, in current test theory and allied procedures, is not the assumption of «something in the heads» of testees that correlates with their test behavior. Rather, it is that the evolution of testing, as we see it today, has been almost as if it had been designed to obscure these «somethings» and to be as unnatural as possible. By «unnatural» I mean here primarily the static conception of human qualities. Certainly, we must have invariant qualities in order to measure. If *all* were really Heraclitian change, then no properties of the change could be described.

Yet, consider the possible fruits that a more dynamic account of people on the part of the test theorists and test makers, might bear. Such an account could lead to a healthy tendency to search for those aspects, or parameters of psychological processes, that are humanly meaningful, dynamic, and useful in predicting future behavior and the potencial for certain classes of behavior.

Certainly, the above program seems in the spirit of such investigators as Susan Embretson (In press) who consciously seeks to build test procedures based on process models of cognition and of course much of the work of Robert Sternberg seems to fall fairly into this domain (e.g., 1977). I believe we are just seeing the incipient and somewhat inchoate roots of what will ultimately be a major stream in psychology, with much cross-fertilization between the test theorists and laboratory cognitive psychologists. Probably psychobiology and perhaps behavioral genetics will also provide useful information. As a psychologist with strong leanings toward mathematical modeling, I have a further bias toward the thesis that putting hypotheses about cognitive processes in mathematically based models, will prove of considerable aid in advancing the cause of a true synthesis of cognitive theory and test theory. (Perhaps some, such as Embretson, might suggest that such aid was already in progress.)

In summing up, I will hazard a possibly inflammatory analogy. There is a branch of applied mathematics and engineering addressed to identification of the internal structure and also the time-dynamics of «machines»; the latter interpreted in the broad sense (e.g., Booth, 1967; Padulo and Arbib, 1974). With some care and a little risk, ordinary psychologists may view themselves as engaged in a similar, if vastly more complex project: to determine a system, or canonical class of systems, that minimally support the observed input-output behavior of their subjects. Test makers, however, must at the present time, rely very heavily on the perspicacity and excellent training of the test readers, interpreters and assessors. It therefore becomes a very real possibility that these tasks or better, the behaviors (and behavors!) involved, should be a part of the overall theory of the test process. At this, juncture, then, our overall approach or metatheory brings in the concerns expressed by Dr. French in the remainder of his tratise, in a quite natural and felicitous manner.

References

BOOTH, T. L. (1967): *Sequential Machines and Automata Theory*, New York, Wiley Press.

EMBRETSON, S. (in press): «Diagnostic testing by measuring learning processes: Psychometric consideration for dynamic testing». In Frederiksen, N. and Lesgold A. (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition*, Hillsdale, N. J.: Earlbaum Assoc.

GAITO, J. (1980): «Measurement scales and statistics: Resurgence of an old misconception», *Psychological Bulletin*, 87, 564-567.

KENDALL, M. G., and STUART, A. (1973): *The Advanced Theory of Statistics* (Vol. 2), New York, Haffner Publishing Co.

PADULO, L., and ARBIB, M. A. (1974): *System Theory*, Washington, D. C., Hemisphere Publishing Co.

STERNBERG, R. J. (1977): *Intelligence, information processing and the principles of cognition*, Hillsdale, N. J.: Erlbaum Assoc.

TOWNSEND, J. T., and ASHBY, F. G. (1978): «Methods of modeling capacity in simple processing systems». In J. Castellan and F. Restle (Eds.), *Cognitive Theory* (Vol. 3), Hillsdale, N. J., Erlbaum Assoc.

TOWNSEND, J. T., and ASHBY, F. C. (1983): *The Stochastic Modeling of Elementary Psychological Processes,* Cambridge, United Kingdom, Cambridge University Press.

TOWNSEND, J. T., and ASHBY, F. G. (1984): «Measurement scales and statistics: The misconception misconceived», *Psychological Bulletin,* 96, 394-401.

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 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 on believes that with some sutiable «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 subjects 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 betweeen 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 thinck 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 lor 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, the 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?