J. R. Statist. Soc. A (1994)
157, *Part* 3, *pp.* 317–356

Deconstructing Statistical Questions

By DAVID J. HAND†

The Open University, Milton Keynes, UK

[*Read before* The Royal Statistical Society *on Wednesday, December 15th, 1993, the President,* Professor D. J. Bartholomew, *in the Chair*]

SUMMARY

Too much current statistical work takes a superficial view of the client's research question, adopting techniques which have a solid history, a sound mathematical basis or readily available software, but without considering in depth whether the questions being answered are in fact those which should be asked. Examples, some familiar and others less so, are given to illustrate this assertion. It is clear that establishing the mapping from the client's domain to a statistical question is one of the most difficult parts of a statistical analysis. It is a part in which the responsibility is shared by both client and statistician. A plea is made for more research effort to go in this direction and some suggestions are made for ways to tackle the problem.

Keywords: METASTATISTICS; THEORY

1. INTRODUCTION

This paper asserts, and attempts to illustrate by a series of examples, that much statistical analysis and design is misdirected. Instead of sorting out precisely what question researchers need to ask, and then finding ways to answer those questions, many statisticians are guilty of pursuing mathematically tractable alternatives which are potentially misleading to the researcher. This is not merely an issue of using simplifying models because the full reality is too complex to cope with—some of the examples given below can be correctly tackled by standard methods. It is a question of ambiguously or incorrectly stated research aims—of making 'errors of the third kind' (giving the right answer to the wrong question).

The aim of this paper is to stimulate debate about the need to formulate research questions sufficiently precisely that they may be unambiguously and correctly matched with statistical techniques. I call this exercise 'deconstruction'. (I have deliberately avoided the word 'analysis' as this is used to describe a particular level of statistical operation—the application of statistical tools to identify structure and patterns in data. This paper addresses a higher level issue—the level which determines what the questions are in the first place and which tools should be used.)

The paper may be regarded as being about *metastatistics* rather than statistics *per se*, in that it is not concerned with narrow aspects of statistical inference or the mathematics of techniques. Or it may be regarded as being concerned with theoretical aspects of statistics distinct from those typically considered in 'methodological' journals.

†*Address for correspondence*: Faculty of Mathematics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK.
E-mail: d.j.hand@uk.ac.open

0035-9238/94/157317

The term *statistical strategy* may also be introduced in this context. This has been used by various researchers in the past (for example, Cox (1977), Cox and Snell (1981), Chatfield (1988, 1991), Hand (1986, 1990a) and Oldford and Peters (1986)) to describe issues of when particular techniques should be applied, how to use them and how to interpret the results, as distinct from issues concerned with lower level aspects of implementing given methods. Chatfield (1991) has remarked that statistical texts tend to concentrate on technique at the expense of strategy, and yet understanding what experienced statisticians would look for in the data and how they would undertake their analyses is just as important. This paper may be regarded as being concerned with an early stage of a statistical strategy: clarifying the questions that the researcher wishes to consider.

The structure of the paper is as follows. Section 2 describes some of the important issues to be considered when attempting to deconstruct a research question. When expressed in general terms some of these may seem obvious, even trivial. And yet, as the examples show, the fact that they may be obvious when articulated has not prevented mistakes from being made.

Section 3 presents a series of examples. Some of these will be well known to experienced statisticians so that a problem seldom arises in practice. Others will also be well known—but nevertheless often give rise to misunderstandings and, presumably, incorrect conclusions. Yet other examples are less well known. The issues raised in the examples are widespread, and although particular papers have been cited the aim is not to criticize the authors but simply to demonstrate that the problems illustrated are genuine and not unrealistic artificial constructs.

Section 4 examines the material from the reverse viewpoint: instead of focusing on the question with the aim being to clarify it and hence to see how statistical methods can be used to address it, statistical techniques are considered, with the aim being to see precisely what questions they answer.

I remarked earlier that the primary aim of this paper is to encourage discussion about the issues involved in formulating research questions and deconstructing statistical questions, with the ultimate goal being to minimize ambiguity in question formulation. A significant step towards this goal would be the statement of a number of principles which researchers and statisticians could follow. It is clear that constructing such principles is no small task and requires careful thought. However, by way of stimulus, Section 5 presents some proposals.

2. ASPECTS OF DECONSTRUCTION

It has been suggested that, to become a competent statistician in any particular application domain, one needs at least three years working in that domain after taking a degree in statistics. The implication of this is that there is more to statistics than merely the mathematics of the techniques. It also suggests that many current statistics courses are not teaching some aspects of what is required to become an effective consultant, and that these aspects concern how to formulate and identify, in the first place, the question which is to be answered by using statistics.

This point, that there is more to statistics than merely the choice and application of a set of techniques, has been discussed by others, including Cox (1977) and Kempthorne (1980). In the present paper I am concerned with identifying precisely (in so far as this is possible) what it is the researcher wants to know—an aspect of statistics which precedes the choice and application of techniques (though, inevitably, there will be some mixing of the two—see later).

Statistical studies come in many types. A common distinction is between hypothesis generation and hypothesis testing studies. The former are much less formal than the latter, in that they are relatively unconstrained by limitations on what it is 'permissible' to do. In principle any technique can be used which may reveal some structure in the data, on the basis that such revealed structure is only suggestive, and will be the subject of a later hypothesis testing study. In practice, however, some approaches are much more likely to produce spurious structures than others and so some moderation is necessary. It is also typically necessary to decide beforehand what *sort* of structures might be of interest. Without some kind of definition, however informal, the exercise becomes pointless—any arbitrary distribution of data points is, after all, a 'structure'. Thus, in projection pursuit, for example, an 'interestingness' measure is defined. At the other extreme, different members of the class of exploratory techniques called cluster analysis use different definitions of 'cluster' and all too often too little attention is given before an analysis to what sort of structure would qualify in a particular context as a cluster. The point is that even for informal hypothesis generation studies it is necessary to consider in detail beforehand what it is we want to know.

Hypothesis testing studies involve collecting data and studying how well those data match the theory to be tested. Naturally we wish to make the test as sensitive as possible. That requires collecting the data so that slight departures from the theory manifest themselves clearly in the analysis. In turn that means *designing* the data collection exercise from the perspective of the question to be answered. Again, then, to design a study effectively, it is necessary to have a clear idea of the research question.

Often (typically?), of course, studies are mixed. They often include a hypothesis testing component for which a particular question is posed in advance, and then a hypothesis generating component in which the data are examined to see whether there are any other interesting patterns in them.

The generation–testing distinction is important. Another is the mechanism–description distinction between model types. Much of statistics is concerned with fitting a model to data. In some situations this will be an explicit attempt to model the mechanism through which the data were generated but often the model is simply a summary or description of the data. An example of summarizing the data might be the use of regression in psychology where it is being used simply to show an approximately linear increasing relationship between two measured variables without there being any detailed underlying scientific theory. In general a mechanistic model will carry more contextual baggage which needs to be kept in mind when considering appropriate statistical tools. However, by definition, a mechanistic model tells us what sort of shapes and structures we are looking for. A descriptive model, in contrast, is much less constrained and so requires more care in deciding what sorts of things may be valuable.

Model fitting involves optimizing some criterion. Some criteria have attractive theoretical properties, but all too often the criterion is adopted by default with no thought being given to its suitability to the problem at hand. Modern computer power, however, has opened up the feasibility of using any of a vast range of criteria. Different criteria have different properties and it is necessary to consider which one best matches the aims of the study. This comment applies to simple models, such as location measures, as well as to complex models.

What research questions it is possible to ask will depend on the data. This means that it is necessary to plan what data to collect on the basis of what it is we want to know. The term 'data' includes the numerical values recorded and also the meaning of the numbers and the variables—the context in which the data arose, or the 'metadata' describing the numbers. It is probably impossible to create a general theory of context: the essence of the term is that it depends on the particular application. (Lunneborg (1992) gives the nice example of a confidence interval for the regression coefficient of income on education in a salary survey. An interval of $[-10, 8]$ (dollars per year, say) tells us that the coefficient is effectively 0. An interval of $[-3000, 4000]$ tells us that we have no idea what the slope is. These different interpretations can only be made given the context of the interval.) However, there are aspects of context which are general and about which theories can be constructed. One such is measurement scale, discussed below (see, also, Hand (1993a, b)). Measurement scale can constrain the sort of questions which it is sensible to ask of a particular set of data—and hence can limit the scientific questions which can be posed. Other simple general examples of how the context in which the numbers arose can influence the analysis are when there are bounds on the numbers, when zeros are structural, when they are not a random sample, when they are or are not matched, when they are hierarchical or when data are non-ignorably missing.

Considerations arising from these and other contextual issues will influence the questions which it is meaningful to ask and, at the next level down, the choice of statistical tools.

3. EXAMPLES OF INAPPROPRIATE METHODS

3.1. *Example 1: Explanatory versus Pragmatic Studies*

The first example will be familiar to medical statisticians, though it also applies widely outside the medical field. Schwartz *et al.* (1980) call it the *explanatory versus pragmatic* distinction. They illustrate with a comparison between two radiotherapy treatments (in a standard two groups randomized clinical trial):

(a) treatment group 1—radiotherapy alone;
(b) treatment group 2—radiotherapy preceded for 30 days by a sensitizing drug.

Consider two designs, D1 and D2. In D1 radiotherapy is started immediately for the first group. This is what would happen in a clinical setting. In design D2, in contrast, the radiotherapy alone group first wait through a 30-day period in which no treatment is given. This permits a comparison of two groups for whom the only difference is the fact that one received the drug.

D1 here is *pragmatic* because it is what would happen in clinical practice. In contrast, D2 is *explanatory* because it is seeking to understand, or explain, the effect of the drug, all other things being equal. Both questions are legitimate—neither is 'right' or 'wrong'. Which is appropriate depends on what the investigator wishes to know. And it is easy to show that an answer to one question may not help in answering the other.

In this example it was necessary to decide beforehand which question was to be answered so that an appropriate design could be adopted, and the distinction is, in fact, pervasive. A pragmatic trial must guard against identifying the poorer treatment as better. (Schwartz *et al.* (1980) term this a type III error. This is different from

TABLE 1
Two treatments for ulcers: (a) explanatory analysis and (b) pragmatic analysis

	(a)		*(b)*	
	Drug 1	*Drug 2*	*Drug 1*	*Drug 2*
Recovered	11	13	11	13
Not recovered	7	2	18	13

the errors of the third kind defined in Section 1.) Moreover, type I errors do not matter: if the two treatments are equally effective (this is what the null hypothesis says) then it does not matter whether we conclude that A is better or that B is better. This means that the α-level (the probability of concluding that there is a difference when none exists) can be set at 100%. In contrast, in an explanatory trial we will wish to avoid type I errors and type II errors. This is a classical hypothesis testing situation. Since the two types of trial focus on different aspects of the error structure, it will be apparent that the number of subjects needed will also be influenced by the explanatory–pragmatic distinction. Pragmatic questions correspond roughly with intention-to-treat analyses, whereas explanatory questions correspond more to treatment-received analyses.

Table 1 shows results from a clinical trial of antibiotics in controlling stomach ulcers, giving the numbers who have or have not recovered under each of two treatments after 4 weeks. In part (a) an explanatory analysis is presented. In part (b) a pragmatic analysis is presented, in which patients who dropped out because of side-effects are included and regarded as treatment failures. Although in this case 11 in each treatment group withdrew, the numbers could easily have been quite different and it is clear that in such situations the two approaches may give different results.

Similar points apply to refusals. Carpenter and Emery (1977) described a study of sudden infant death syndrome in which the children in one of the groups were subjected to increased surveillance by health visitors. Among the 627 families who agreed to this there were two unexpected deaths (0.32%) whereas among the 210 who refused there were three unexpected deaths (1.43%). Whether or not the refusals should be included in the analysis, which compared this group with a control group which had no extra surveillance (and 9/922 = 0.98% unexpected deaths) depends on precisely what one is seeking to find out.

Problems like these are particularly common with longitudinal data, where drop-outs frequently occur. Whether the analysis should be based solely on those remaining in the study or be part of a larger model which takes account of the probability of dropping out (see, for example, Diggle and Kenward (1994)) depends on what we want to know. In particular it depends on whether we wish to make inferences to the population of people who do not drop out or to the larger population from which the sample was drawn.

Schwartz *et al.* (1980) pointed out that the explanatory–pragmatic distinction also influences more fundamental aspects of design. For example, in an explanatory study, homogeneous groups of patients will probably be chosen, whereas in a pragmatic study we will choose patients who are representative of the potential treatment populations. This might lead us to suggest that randomization approaches are appropriate in explanatory trials and sampling theory approaches appropriate in pragmatic trials.

In general, it is necessary to look very carefully at the scientific hypothesis that is to be studied to make certain that the correct statistical hypothesis has been formulated and tested.

3.2. *Example 2: Wilcoxon Test*

For our second example we stick to the issue of comparing two treatments. A crude statement of the objectives of such a study might be that we wished to see which of the treatments (A or B, say) was better. However, this statement is inadequate, as we see when we begin to deconstruct it. (Hand (1992) discusses this example in more detail.)

The first thing to ask is what is meant by 'better'? Here we shall adopt the straightforward view that better means a 'larger' score on some measured scale, but even this is insufficient. Individuals have scores; groups do not. Thus we can compare the scores of two subjects in the study, but a more careful definition is required before we can compare two groups.

Pursuing our policy of deconstruction, we might decide that our real interest is in the comparative likely effect of the two treatments on a future patient. This is a common aim for many such studies, i.e. the question that we really wish to answer is 'Will a future subject score higher under A or B?'. If the score of a subject receiving treatment i is denoted z_i then the above question can be expressed as 'Will $z_A - z_B$ be greater than 0?'. Since different subjects differ, the best that we can hope to achieve is an answer to the question 'Is $P(z_A - z_B > 0) > \frac{1}{2}$?'.

To shed light on this question, we have a sample of subjects, each of which we shall assume has been randomly allocated to receive either A or B and we shall further assume that all covariates which might refine the distribution of response errors have been taken into account.

If both A and B can be given to each subject, with no carry-over effects, then we can observe $z_A - z_B$ directly for each patient in our sample and hence estimate the probability that $z_A - z_B > 0$ and test whether it is greater than $\frac{1}{2}$. However, in many situations it is not possible to administer both A and B (treatments influencing survival time until death or methods for teaching reading skills, for example). Matching subjects can side-step the problem, but in many treatment comparisons the subjects are not matched: the comparisons are based on two independent groups. Hand (1992) showed that the Wilcoxon test for two independent samples is equivalent to testing whether $P(x_A - y_B > 0)$ is greater than $\frac{1}{2}$, where x_A and y_B are *independent* scores under the two treatments, i.e. the common design, using the Wilcoxon test for analysis, considers $P(x_A - y_B > 0)$ whereas we really want to consider $P(z_A - z_B > 0)$. Hand (1992) gave examples which showed that one of these probabilities can be greater than $\frac{1}{2}$ while the other is less than $\frac{1}{2}$ (and vice versa). That means that the Wilcoxon test can conclude that treatment A is more effective than treatment B (in the sense described above) when in fact the converse is the case.

Although the Wilcoxon test, in a sense, comes closest to answering the question that we are interested in (is $P(z_A - z_B > 0) > \frac{1}{2}$?), the t-test is often used to compare two groups. However, the t-test is not simply concerned with the sign of a difference between two treatments but also takes into account the size of the difference. Sometimes, of course, this is the right thing to do. It depends on precisely what the researcher wants to know—i.e. on a careful deconstruction of the research objective.

3.3. *Example 3: Proxy Variables*

Measurement theory is the discipline concerned with establishing and studying the properties of mappings between objects being studied and the numerical (or otherwise) representation of them (see Section 4.1). Some areas of this discipline have been very well developed, and extremely powerful results have been established. However, just as with probability and statistics, different schools exist and not all issues have been resolved. Sometimes insufficiently clearly formulated research questions arise from such issues. This example illustrates such a case where precisely what is being studied has not been sufficiently clarified. The numbers have been deliberately chosen to keep the example straightforward, and more realistic but complicated numbers could be chosen.

Consider two researchers, whom, to lend force to the illustration, we shall take as a Frenchman and an Englishman, who wish to compare the fuel efficiencies of two types of car. To do this they take samples of two cars of each type and measure the fuel efficiency.

The results are shown in Table 2.

The English researcher records the data in the first row of the table, finding that the two cars of type 1 ran for 1 mile per gallon and 4 miles per gallon respectively, producing an average of 2.5 miles per gallon. Similarly for type 2 cars this researcher found that the average efficiency was 2.0 miles per gallon. The English researcher thus concludes that type 1 cars are better—running for more miles per gallon than type 2 cars.

The French, however, measure fuel efficiency in the metric equivalent of gallons per mile. The French researcher thus uses the reciprocals of the English researcher's raw scores. For car type 1 this means that the two cars consume 1 gallon per mile and 0.25 gallons per mile respectively, producing an average of 0.625 gallons per mile. And for car type 2 the average is 0.5 gallons per mile. The French researcher thus concludes that type 2 cars are better—consuming less petrol per mile than type 1 cars.

We appear to have a contradiction. How can the two researchers reach opposite conclusions with the same data?

The difference between the two conclusions clearly arises because of the non-linearity of the reciprocal transform. In general, if f is non-linear, then $E\{f(x)\}\neq f\{E(x)\}$.

This is all very well, but it does not answer the real question: which researcher is correct? (Or, operationally, which car should I buy?) Since there are two distinct answers, which are based on the same data, it seems that the question has not

TABLE 2
Fuel efficiencies of two types of car

Researcher	*Efficiencies for type 1*			*Efficiencies for type 2*		
	Car 1	*Car 2*	*Average*	*Car 1*	*Car 2*	*Average*
English (miles per gallon)	1	4	2.5	2	2	2.0
French (gallons per mile)	1	0.25	0.625	0.5	0.5	0.5

been sufficiently carefully phrased—the structure of the question needs to be elucidated.

In this particular example we might argue that, in using a car, we are generally interested in how many gallons it will take to cover a given distance (to travel from A to B) rather than how far we can travel on x gallons, before we run out of petrol. That being the case, the gallons per mile calculation will be the more appropriate (with the implication that the English are wrong!). It is the context, as noted in Section 2, which here suggests that this is the appropriate formulation of the question, and different problems will have different contexts, yielding different questions and different solutions. Indeed, in problems involving ratios, such as this problem, the context may suggest yet a third alternative: not an average of ratios or an average of their reciprocals, but rather a ratio of their averages.

Or, yet another alternative, for the car fuel example we might instead assert that we are really interested in 'efficiency', which lacks an adequate operational definition and that we therefore instead use either miles per gallon or gallons per mile as proxy variables which have ready operational meanings. We assume that they are monotonically related to efficiency (they are monotonically related to each other, which is reassuring). In this case, given the weak relationship between the variable of interest and its measurable proxies, if we wish to make a meaningful statement about efficiency we must use techniques which are invariant to monotonic transformations. The arithmetic mean is not, of course, and is therefore an inappropriate tool to use.

In any case, the precise question that the researcher wants to answer must be made clear.

3.4. *Example 4: Lord's Paradox*

'Lord's paradox' (Lord, 1967) is a well-known but nevertheless effective demonstration of the importance of being precise about the research objectives. Lord described a (fictitious) study to explore the 'effects on the students of the diet provided in the university dining halls and any sex difference in these effects'. To investigate this, the weight of each student is measured in September and again the following June. Now, it is clear that, to study the diet effect over this period, we need to examine the final weights, allowing for the differences in initial weights. One approach is simply to study the change in weight over the period. In the situation that Lord describes, the mean change for boys is 0 and the mean change for girls is 0, so this approach leads to the conclusion that there is no sex difference.

An alternative way of examining the final weights allowing for the differences in initial weights is to conduct an analysis of covariance. This leads to identical slopes of final on initial weight in the two sex groups, but different intercepts, with that for the boys being larger than that for the girls. The conclusion from this analysis is thus (Lord (1967), p. 305) 'boys showed significantly more gain in weight than girls when proper allowance is made for differences in initial weight between the two sexes'.

Many researchers have discussed this apparent paradox, and agreed that the fundamental problem is the imprecise statement of the hypothesis being tested. The first analysis above is a test of an unconditional comparison between the gains of the two groups. The second, however, is a test of an average conditional comparison, conditioning on initial weight, i.e. if we select subgroups of boys and girls with identical initial weights, then the relative position of the regression lines shows that the boys gain more weight than the girls. Given that the questions are different, it is hardly surprising that the answers differ.

This is all very well, but, as in so many of the examples given here, this deconstruction does not completely resolve the issue. It is still necessary for the researcher to be clear about which problem should be solved.

Wainer (1991), based on Holland and Rubin (1983), described an extension to Lord's paradox in which the aim was to compare the effects of a treatment on two groups, making due allowance for differing base-line values. Similar points as previously apply to whether we should subtract or covary out the base-line scores, but now an additional complication is that we really wish to compare the post-treatment scores with the scores that the subjects would have had, had there been no treatment intervention. This means that extra (intrinsically untestable) assumptions must be made relating base-line scores with the scores that the subjects would have had later.

To summarize, it is necessary

(a) to deconstruct the statistical question to identify precisely what it is answering,
(b) to know what question the researcher wishes to answer and
(c) to be aware of any latent assumptions so that they can be assessed for realism and, perhaps, so that evidence supporting them can be obtained.

3.5. *Example 5: Simpson's Paradox*

Simpson's (or Yule's) paradox (Simpson (1951) and several references in Haunsperger and Saari (1991)) also falls into the category of well-known but convincing illustrations of the need to construct the research questions very carefully in view of the objectives. Consider the data in Table 3 (Early and Nicholas, 1977) on how the proportion of males in a particular psychiatric hospital changes over time. In 1970 the proportion was 0.464 and in 1975 it was 0.462: a small reduction. However, suppose that we examine the under 65 and 65 and over age groups separately. For the former

TABLE 3
Proportions of males in a psychiatric hospital

	Proportions in the following years:		
	1970		*1975*
All ages			
Male	343		238
——	—	=0.464>0.462=	—
Total	739		515
<65 age group			
Male	255		156
——	—	=0.594<0.605=	—
Total	429		258
⩾65 age group			
Male	88		82
——	—	=0.284<0.319=	—
Total	310		257

group we find the 1970 proportion to be 0.594 and the 1975 proportion to be 0.605: an increase. And for the latter group we find the 1970 proprtion to be 0.284 whereas the 1975 proportion was 0.319: also an increase.

The two age groups separately each show an increase whereas the complete population, which is the combination of the two age groups, shows a decrease. How can this be? Surely it is a contradiction.

It is not, of course, and a simple explanation is as follows. Let $x \equiv$ male, $y \equiv$ under 65, and $z \equiv 1970$, with x', y' and z' being the complementary categories. Then elementary probability theory tells us that

$$P(x|z) = P(x|y, z)\, P(y|z) + P(x|y', z)\, P(y'|z)$$

and

$$P(x|z') = P(x|y, z')\, P(y|z') + P(x|y', z')\, P(y'|z')$$

with $P(y|z) = 0.581$ not equal to $P(y|z') = 0.501$ and $P(y'|z) = 0.419$ not equal to $P(y'|z') = 0.499$.

The consequence is that, although in the example $P(x|y, z) < P(x|y, z')$ and $P(x|y', z) < P(x|y', z')$, the different sets of weights mean that the weighted average of $P(x|y, z)$ and $P(x|y', z)$ is greater than the weighted average of $P(x|y, z')$ and $P(x|y', z')$.

Intuitively, when we average the two age groups, we tend to use equal weights and since both age groups have proportions changing in the same direction their equally weighted average does also.

This is a nice neat mathematical explanation of the apparent paradox, but again I think that it is an inadequate resolution. What is needed is a more careful examination of what the researcher really wants to know. Either method of averaging may be appropriate. The weighted average (overall calculation) is appropriate to answer the question 'Does the proportion who are male increase?' but the unweighted average (calculating the age groups separately and then averaging them) is appropriate to answer the question 'On average, for patients of a given age, does the proportion who are male increase?'. One question is concerned with the change of averages and the other with the average of changes. Which is appropriate depends on a sensitive deconstruction of the research objectives.

In this example, as in example 4, the issue is whether or not the researcher needs to ask a conditional question and, in fact, the issue of whether or not to condition is ubiquitous. It occurs, for example, in simple *versus* multiple regression. Neither analysis is right and the other wrong—it depends on what we want to find out.

3.6. *Example 6: Interaction*

A nice example of the care needed in formulating research questions arose from a controversy in psychiatric research (Everitt and Smith, 1979). The data in Table 4 (from Brown and Harris (1978)) show, for a sample of women from south-east London, the numbers developing and not developing depression over a 1-year period (y) cross-classified by the presence or absence of intimacy with a husband or boyfriend (x_1) and whether or not they experienced a severe life event in this period (x_2).

TABLE 4
Depression and no depression frequencies classified by presence or absence of intimacy (x_1) and experience or not of a severe life event (x_2)

y	x_1			
	No intimacy		Intimacy	
	x_2		x_2	
	Event	No event	Event	No event
Depression	24	2	10	2
No depression	52	60	78	191

Brown and Harris (1978) concluded from these data that the two predictor variables (x_1 and x_2) interact in their effect on the response variable y. Tennant and Bebbington (1978), however, using the same data, concluded that x_1 and x_2 are independent in their effect on y—i.e. that x_1 and x_2 have no interaction. Again we are presented with the question: which team is correct? Again the difference does not arise because of a different choice of significance level, but is caused by a deeper structural difference in the approaches used by the two teams.

Neither team has defined sufficiently precisely what it wants to know. Although both have used the term 'interaction' neither has looked closely at its definition and related it to what it wants to know: they have not deconstructed their questions.

Brown and Harris (1978) used an additive model and defined interaction accordingly. Thus, as Table 5 shows (based on Table 4), the difference between columns in the upper row is 0.29 whereas that between columns in the lower row is 0.10, a substantial difference. This shows (they say) a clear interaction between the two cross-classifying variables.

In contrast, Tennant and Bebbington (1978) used a multiplicative model, with interaction appropriately defined. Now Table 5 shows the ratio of the two cells in the upper row to be 10.7 whereas that in the lower row is 11.0. These are almost identical, showing, Tennant and Bebbington claimed, that the two cross-classifying variables do not interact in their effect on y.

The point of this example is that the two approaches have led to different conclusions. Given that both analyses are conducted, this difference will alert us to the fact that more careful thought is required. But 'both analyses' will not normally

TABLE 5
Proportions with depression calculated from Table 4

x_1	x_2	
	Event	No event
No intimacy	0.32	0.03
Intimacy	0.11	0.01

be conducted. The danger is that inadequate prior consideration of exactly what question was being asked could have led to a mistaken conclusion. Presumably it did for one team in this example, given that they wanted to answer the same scientific question.

3.7. *Example 7: Trimmed Means*

Let $Q(X)$ represent an arbitrary real functional on the distribution of the random variable X. For example, we shall be interested in the functionals $E(X)$, the expectation, $R(X)=P(X>0)$, and $T(X)$, the trimmed mean of X.

Let (X, Y) be a pair of possibly dependent random variables and let $U=X-Y$. Let (X', Y') be independent random variables with X' having the marginal distribution of X and Y' that of Y and let $V=X'-Y'$. We can then ask which functionals Q satisfy

(a) $Q(U)=Q(X)-Q(Y)$,
(b) $Q(U)=Q(V)$?

And, in the spirit of this paper, having identified the functionals with these properties, we could then go on to ask what substantive research questions these functionals are concerned with. However, we shall here focus on some special cases.

In example 2 earlier I pointed out that the Wilcoxon test and the t-test answer different questions. One just considers the signs of differences whereas the other also takes the sizes into account. In that example we wanted to ask a question about the distribution of the z_A-z_B differences, but we were bound by the nature of the experiment to ask about the distribution of differences x_A-y_B from two independent samples. The Wilcoxon test addresses the issue of whether or not $R(x_A-y_B)>\frac{1}{2}$. In contrast the t-test looks at $E(x_A-y_B)$ and compares it with 0. A problem arose because the functional used in the Wilcoxon test does not satisfy (a) and (b) above, so that it is not true that $R(z_A-z_B)=R(x_A-y_B)$ or that $R(z_A-z_B)=R(z_A)-R(z_B)$. Expectation, however, does satisfy (a) and (b)—so, if we had been interested in a comparison taking effect sizes (and not just signs) into account, no problem would have arisen.

Of course, there may be particular distributional forms—additional assumptions which can be made—for which (a) and (b) are satisfied for some Q, even though they are not generally true. We return to this sort of possibility in Section 4.2.

Efron (1992) used jackknife-after-bootstrap approaches to explore the extent of trimming for a problem involving trimmed means. The situation is as follows. A subatomic particle called the tau particle decays soon after production into various collections of other particles. Some of the time the decay produces just one charged particle, and this can happen in four major and various minor ways. The proportion of time that a single charged particle is produced can be estimated and similarly the proportion of time that each of the four major paths is followed can be estimated. The primary interest in the study is to estimate and find a confidence interval for $d=d_0-d_1-d_2-d_3-d_4$, where we have used d_0 to denote the proportion of experiments producing a single charged particle (of whatever kind) and d_1–d_4 to denote the proportion of experiments producing particles of types 1–4.

The purpose of Efron's investigation was to show how the jackknife-after-bootstrap method can be used to compare a series of trimmed means estimators, permitting one to choose that producing the most accurate estimate. This question seems to refer to trimmed mean estimators of d. A problem arises, however (Efron (1992), pages 95–96): 'because of certain physical constraints, any one experiment provides only one estimate . . ., either an estimate for the composite rate decay, or for one of the four modes'. Because of this, instead of using $T(d)$ (with T denoting the trimmed mean), Efron used $T(d_0) - T(d_1) - T(d_2) - T(d_3) - T(d_4)$. However, in general T does not satisfy (a). It follows that the two estimators may produce different results.

The goal was to produce a confidence interval for the difference parameter d. To do this, we need to study the distribution of the differences. What has been done, however, is to study the difference of the distributions (and in particular the difference of the trimmed means of the distributions). The two need not give the same results—they are answering different questions.

3.8. *Example 8: Sums of Squares with Unbalanced Data*

I imagine that, once the issues have been clarified, most statisticians would reach agreement on most of the previous examples. But this is not always the case. Take the example of a two-factor analysis of variance with unbalanced data. The SAS GLM routine gives a choice of four different types of sums of squares for such an analysis, denoted types I, II, III and IV. This inevitably raises the question of whether researchers appreciate the distinctions between the types and can match one of the four with their research question. But deeper issues are also involved in this example. Presumably SAS proponents believe that each of these four types of sums of squares is valuable. Nelder (1992), however, argued that the SAS literature is confused about the distinction between the hypothesis to be tested and the non-centrality parameter in the expectation of the numerator sum of squares of the F-statistic. As a consequence, he asserts (Nelder (1992), p. 403): 'Type III and Type IV sums of squares serve no useful inferential purpose and should be abandoned'. He goes on to point out that type III sums of squares break marginality requirements. Hypotheses ignoring marginality requirements are, he asserts (Nelder (1982), p. 142), 'without practical interest'—and yet users of SAS GLM are presumably testing such hypotheses.

The situation is aggravated when one steps beyond SAS and also considers other packages. As Searle (1987) says:

> 'As a result of having several methods of analysis, not all statistical computing packages necessarily do the same analysis on any given set of unbalanced data. Consequently, in the context of hypothesis testing, or of arraying sums of squares in an analysis of variance format, there is often, for the one data set, a variety of sums of squares available from computing packages. The problem is to identify those that are useful.'

Or, as I would say, the problem is to identify those which address the hypothesis that one is interested in.

In an elegant discussion of the historical background, Herr (1986) raised the role that computational aspects played in the derivation of, and attempts to choose between, the different sums of squares. Finney (1948) is cited as making the point that the definition of main effects should depend on what the analysis is all about and that 'elegance of analysis alone must not be the criterion'. By way of conclusion, Herr wonders whether, had the early researchers stated their hypotheses clearly, might not their papers, instead of leading to decades of discussion of computational complexities, have

> 'sparked a series of papers arguing when to use each analysis? And based on more precise information about what the different methods tested might not these arguments have been more fruitful? The 21 years since 1965 might well have been better spent had this been the case'.

In a sense, this entire paper is trying to make the same point more generally: not just relative to computational aspects, but to general issues of what questions statistical techniques are concerned with. And, rather than wondering about how past effort might have been better spent, it is an attempt to stimulate the more fruitful use of our future efforts.

4. USE OF DIFFERENT METHODS AS ALTERNATIVES

In Section 3 I approached things from the perspective of the presenting question, showing how important it was to formulate that question precisely. This, as I have been at pains to point out, is the correct way to think about research questions: scientists have particular issues that they want to resolve, so that the analysis must begin by focusing on those issues. Nevertheless, when addressing a statistical audience it is also useful to discuss things from the other perspective, that of the statistical technique which might be applied. This is the perspective that I adopt in this section, again showing that insufficient care is often exercised when choosing techniques.

At this point it is helpful to make precise some terminology which is normally used fairly loosely. For me, here, a model will be a *family* of mathematical descriptions rather than a particular member of a family. Thus a model requires the values of parameters to be provided to give a complete specification. This means, for example, that we can talk of a linear regression model, and that the parameters of this model can be determined by a variety of techniques according to the criterion of goodness-of-fit that we choose to optimize (least squares, least absolute deviations, etc.). A statistical test will assume model forms for the populations and then will impose additional restrictions through the null and alternative hypotheses. For example, the model for a Student *t*-test is that the data arise from normal populations with equal variances, and then the null hypothesis imposes the additional structure that the difference between the means is 0. A test thus has two components, a *model* and a *hypothesis*. Both of these play a fundamental role in the choice of the test, but all too often greater emphasis is placed on the model aspects. This can have disastrous consequences for the relevance of the test to the research question being investigated.

To make progress I now need to consider the nature of hypotheses. Without wishing to delve too deeply into the philosophy of science, I shall suppose that the objective of a scientist conducting research is to make some statement about properties of the objects that he or she is investigating, i.e. the scientist starts with a *scientific hypothesis* that they wish to test (that one object is heavier than another or that a group of people are, on average, more intelligent than another group, for example). To apply statistical methods it is first necessary to translate this scientific hypothesis into a *statistical hypothesis*. This translation involves all the usual issues of experimental design, randomization etc., but a key component is deciding how and what to measure. This is the aspect that I focus on in Section 4.1.

4.1. *Hypotheses and Measurement*

A traditional view is that measurement consists of a mapping from the objects being studied to numbers, in such a way that the relationships between the numbers correspond to relationships between the objects. In general, the numerical representation will not be unique—there will be a class of *admissible transformations* describing mappings between alternative legitimate numerical representations. For example, we can choose to measure the weight of an object in grams rather than ounces, with the numbers used to represent the weights in the different measurement systems being related through similarity transformations. The existence of these admissible transformations induces a classification of measurement structures. If the admissible transformations are all one-to-one transformations then the measurement structure is said to have a *nominal* scale. If the admissible transformations are all monotonic increasing transformations then the structure has an *ordinal* scale. Affine transformations lead to *interval* scales. Similarity transformations lead to *ratio* scales.

On this basis, Stevens (1946) argued that, since the objective of science is to draw conclusions about the empirical objects under study, the only statistical operations that were legitimate were those that were invariant under the relevant class of admissible transformations. Other statistical operations would give results which would vary according to the numerical representation adopted and since each representation (obtained by an admissible transformation) was equally valid the conclusions would be valueless. In my terminology, this implies that only certain hypotheses can be meaningfully formulated: others are technically 'meaningless'. For example, the hardness of rocks is measured on the Moh scale, which preserves merely the ordinal relationship between the empirical objects (the rocks). A hypothesis stating that the arithmetic mean value of the hardnesses of one group of rocks was greater than that of another group could give results which differed according to which of the legitimate numerical representations was adopted. Such a hypothesis would be meaningless.

This seems quite straightforward, and it implies that the measurement scale influences the choice of statistical method. But it does this via the research question that is being explored, and not directly.

This view has been adopted by many people. Among these have been enthusiasts for nonparametric tests, arguing that the less restrictive distributional assumptions that these make mean that they may legitimately be applied with weaker (e.g. ordinal) measurement scales. Unfortunately, however, the presentations sometimes lose sight of the fact that the nonparametric tests often achieve their legitimacy only at the cost of changing the hypothesis being tested—from a comparison of means, for example, to one of medians. To put it in the terminology defined above, discussion has focused on the model instead of on the hypothesis. The latter must come first. To do otherwise can result in testing a hypothesis other than that in which the researcher is really interested.

Before leaving this issue, we should note that, although Stevens's restrictions on permissible statistical techniques apply when there is a clear empirical system which is being modelled (as in most of physics, for example), not all science deals with such cases. In particular, test scores and rating scales in psychology do not have clearly defined empirical counterparts to serve as the domain of the mapping to the numbers. When this is the case, it has been argued, Stevens's restrictions do not apply, and the choice of statistical technique is unrestricted by questions of measurement scale. To pursue this would lead us away from our central concern and we shall merely comment here that this issue has been the focus of debate which has rumbled on for most of this century. A detailed historical account of the debate is given in Hand (1993a).

4.2. *An Invariance Principle*

Earlier I defined a 'model' as a family of mathematical descriptions and a hypothesis as a restriction on that family. The model effectively summarizes the assumptions that we are prepared to make about the distributions of the variables under study, whereas the statistical hypothesis, derived from the scientific hypothesis, restricts the distributions yet further. A clear statement of both is important in formulating a statistical test. However, whereas ignoring the model means that the results may not be relied on, ignoring the hypothesis means that we may be testing something which the researchers are not interested in.

Sometimes, however, researchers find it difficult to formulate their objectives precisely. Indeed, any practising statistician will be familiar with this fact. In trying to formulate optimal classification rules, for example, it can be extremely difficult for researchers to articulate the relative costs of different kinds of misclassification. Even in the canonical problem of comparing two groups it is often difficult for researchers to state what comparison is really of interest to them. The choice of wishing to compare two means is all too often made simply because that is what the statistician (or the text-book) suggested. As I have pointed out earlier, this is quite wrong: the choice should have emerged from the scientific aims.

One is tempted to criticize ill-formulated research objectives, but it is perhaps not entirely reasonable to be too proscriptive. After all, every science has shaky foundations if we look sufficiently closely (even statistics, with its disagreements over how to interpret probability). So the research hypothesis itself might legitimately be regarded merely as an approximation to what we really want to know. The problem is that (as illustrated in Section 3) an insufficiently precise formulation may lead to researchers with superficially the same question obtaining different answers, even with the same data.

One possible strategy for tackling this difficulty is to apply a principle of invariance. To illustrate, researchers may be unable to decide whether they wish to compare two groups by using means or by using medians. In general, tests of the two hypotheses (that the means are equal and that the medians are equal) will lead to different results. However, if we are prepared to assume symmetric distributions then the tests are (logically) equivalent. Thus the principle of invariance is: choose the model class such that the truth value of the research hypothesis is (logically) invariant over the different statistical hypotheses which may be used to describe it. At the very least, this makes explicit where the assumptions lie and ensures consistency of conclusions. Moreover it forces thought about just what the hypothesis is.

The equivalence of two hypotheses under a particular model for the population distributions need not imply that the sample-based tests will have the same outcome. But now a choice can be made by applying (any) other criterion—both tests, after all, explore the same hypothesis. Choice could, for example, be based on questions of relative power.

To generalize, suppose that a researcher is unwilling or unable to decide which of two hypotheses H_1 and H_2 she wishes to test. Suppose also that two possible models, M_1 and M_2, are being contemplated, with M_1 more restrictive than M_2 (denoted

$M_1 < M_2$). Then it may be that $H_1 \# H_2|M_2$ but $H_1 = H_2|M_1$ (where # signifies non-equivalence of the hypotheses, = signifies equivalence and | signifies conditioning).

The models form a lattice with the least restrictive models towards the top and the most restrictive towards the bottom. Equality of two hypotheses at some node of the model lattice implies equality at lower nodes. This means that if the researcher can identify some model node such that all hypotheses of possible interest become logically equivalent at this node then there is no need to refine the hypothesis in greater detail.

The invariance principle tells us that two hypotheses are equivalent under certain models M, and the principle is useful if we cannot decide which of the hypotheses we want to test. However, it can also be applied to situations when we know that we want to test H_1 but circumstances mean that we cannot, though we can test H_2, which is related to H_1. What we do is to identify a model M such that $H_1 = H_2|M$. For example, in example 2 earlier let us model the ith subject's z_B-score by $z_B(i) = z_A(i) + s(i)$ and consider two models for $s(i)$:

(a) M_1—$s(i)$ follow arbitrary distributions, possibly different for each subject, but all having medians of the same sign;
(b) M_2—$s(i)$ are arbitrary.

The hypothesis of interest is H_1: $P(z_A > z_B) > \frac{1}{2}$ but, as explained in that example, we cannot test it by using the available data. We can, however, test H_2: $(x_A > y_B) > \frac{1}{2}$. Now, $H_1 \# H_2|M_2$, which is what we would ideally like, but $H_1 = H_2|M_1$. Thus if we are prepared to adopt a more restrictive model—to make additional (untestable) assumptions—then we can use the Wilcoxon test (see Hand (1992)).

We have already noted in example 3 of Section 3 that the two different fuel consumption variables yield the same result if we embed the investigation in a model in which they are simply indicator variables, ordinally related to an underlying latent variable which is the object of real interest. However, this embedding implied that we could not (for example) compare the two car types by using means—the model has restricted the hypotheses which it is meaningful to ask.

Example 8 of Section 3 presented four different types of sums of squares which might be considered when conducting an analysis of variance. Using I, II, III and IV to indicate hypotheses on each of the types, the invariance notation tells us that II = III = IV|(model does not include interactions).

Different hypotheses may also be equivalent when certain data structures hold. Sometimes studies can be deliberately designed with this aim in mind. The SAS analysis of variance sums of squares illustrates this. We have I = II = III = IV|(data are balanced) and III = IV|(no empty cells). We might be inclined to use the former of these to avoid the need to think carefully about hypothesis formulation by using a balanced design (and this would avoid the point raised by Nelder, discussed earlier).

Invariance notions are related to robustness. Let $T_i|M_j$ mean that some technique T_i will produce reliable results under model M_j and let $-T_i|M_j$ mean that technique T_i will not produce reliable results under M_j. ('Reliable' here may mean various things. The arithmetic mean will be an 'unreliable' estimate of the mean of a normal distribution when there are asymmetric contaminating outliers. A trimmed mean will be more reliable.) Then, if $T_1|M_1$ but $-T_1|M_2$ while $T_2|M_2$ and $M_1 < M_2$ then T_2 is said to be more *robust* than T_1. We might, alternatively, say that the behaviour of T_2 is invariant to a broader class of models than is the behaviour of T_1.

In summary, often a scientific hypothesis can be translated into a statistical hypothesis in several ways and it may not be possible to identify which is really intended. When this is the case, it may be possible to adopt a model form under which the competing hypotheses become equivalent. Similarly, we may be unable to test the hypothesis of real interest, but this may be equivalent to a testable hypothesis under certain models.

Of course, it hardly needs saying that the models have to be realistic. A desire to evade thinking about the precise objectives of the research should not seduce us into accepting unreasonable models.

However, in many situations it is known that all effects in some saturated model exist in reality, but some of them are very small. The objective of statistical model building is to identify the larger, more important effects (perhaps so that other effects can be estimated with greater accuracy—Altham (1984)). This means that we are not choosing the model simply on the basis of theory (if we were doing that there would be no need for the statistical analysis) but to achieve descriptive simplicity. This being the case, the adoption of a (realistic) model form which simplifies subsequent analysis seems only sensible.

5. SOME GENERAL PRINCIPLES

The examples in Section 3 were intended to provide evidence that extreme care is needed in formulating research questions and establishing accurate mappings from the scientific to the statistical hypotheses. They were also intended to show that sufficient care is not always taken. It would be of more value, however, if we could go further than simply saying 'one must take care' and explicitly state some general guiding principles of question formulation. Ideally such guidelines need to be sufficiently general to have reasonably broad applicability, and yet sufficiently specific to provide concrete help. They need to be stated in such a way that researchers can see how their problems match the situations described and they need to include tactical advice on what to do in each case. Developing such guiding principles is clearly a major task and one of the objectives of this paper is to stimulate discussion in this area. To set the ball rolling, therefore, some tentative initial suggestions are now presented.

5.1. *Individual versus Populations*

Do we wish to make a statement distinguishing some aspect of the distributions of populations (e.g. comparing the average brain weights of men and women) or do we really wish to make a statement about individuals? This distinction arose in examples 2 and 7 and is related to questions of conditioning on other factors (see Section 5.2). It is the distinction between cross-sectional studies (comparing groups at a particular time) and longitudinal studies. One way in which we might identify which of the two cases we are concerned with is to ask whether interest lies with a 'difference between distributions' or with a 'distribution of differences'.

5.2. *Conditioning*

Simple regression tells us the effect of some variable x on a dependent variable y. Multiple regression tells us the effect of x on y over that which may be attributed to z. The regression coefficients of x can be very different in the two cases and can even differ in sign. However, both analyses are legitimate—it is simply that they are concerned with different questions.

This sort of situation arises in other contexts as well. For example, we may choose to analyse either a cross-classification of three categorical variables or a two-dimensional marginal. Collapsing the third factor loses information about how the other two behave at different levels of this third factor but may be legitimate: it depends what we want to know.

Examples 4 and 5 illustrate problems of this kind. It also arose in example 2 in that the nature of the question really required matching subjects (given that subjects could not receive both treatments).

The issue is conditioning on other variables, controlling for other variables or holding other variables constant.

5.3. *Ambiguous Definitions*

One general area which causes problems is that of ambiguous definition. This leaves open the possibility of alternative operationalizations, so that different researchers may draw different conclusions, and is illustrated in examples 3 and 6. The problem may be eased by a more cautious approach, involving both researcher and statistician, to the translation from scientific to statistical hypotheses, but this will certainly not remove all difficulties. More thought about the nature of the link between what is being studied and the numbers used to represent it can also help.

5.4. *Pragmatic versus Explanatory Questions*

The pragmatic *versus* explanatory distinction is pervasive, and yet seems to defy simple encapsulation. In its simplest form it is the difference between two definitions of the population being studied (e.g. all humans (explanatory) *versus* those who stick to the protocol (pragmatic)). It can also be described as the difference between an exploration of the underlying mechanism (explanatory) and an exploration of the behaviour within a population (pragmatic). This last description shows that the distinction is much more widely applicable than merely to clinical trials.

5.5. *Multiple Univariate versus Intrinsically Multivariate Questions*

Some disciplines, notably psychology, make extensive use of multivariate techniques involving multiple dependent variables—such as multivariate analysis of variance. A natural question then arises about whether the multivariate approach is answering a research question of interest. A useful distinction is between *multiple univariate* problems and *intrinsically multivariate* problems. In the former, interest lies in each variable separately. So, for example, we might be interested in the effect of a treatment on bacterial population, on inflammation and on temperature, with each question being of interest in its own right. This would be a multiple univariate study and the appropriate analysis would be three separate univariate analyses of variance. (See, for example, Hand (1990b).) Conversely, we might be interested in the overall health of two groups, with health measured by a number of items. Here, given that we are interested in an overall result, it would be inappropriate to analyse the items separately. The individual items are of no intrinsic interest—they are merely indicator variables for the question of interest. The situation is intrinsically multivariate and a multivariate analysis of variance using all the items is appropriate.

An explanatory study is likely to be multiple univariate whereas a pragmatic study is more likely to be intrinsically multivariate.

6. CONCLUSION

The aim of this paper is to stimulate discussion about the relationship between scientific and statistical questions. Establishing the mapping from the substantive world of the statistician's client to the mathematical world of the statistical technique is where difficulty lies. In a sense it is more difficult than the mere mathematics of statistical techniques because it is less well formalized. And we might also argue that establishing a valid such mapping is more important than applying rigorous mathematics to the problem formulation which results: it is better to have an approximation (if we know that it is an approximation) to the question that we want to ask, than to have a mathematically correct solution to an irrelevant question.

The paper suggests that perhaps in the past statisticians have not been as careful as they might have been in sticking to the essence of the researcher's problem. To some extent this distortion was justified because of computational difficulties. To do statistics properly, we need to know what the research question is, to have an effective strategy for answering that question, and to have the requisite mathematical or algorithmic facility. Without the last of these, or with inadequate algorithms, no progress can be made, so this last is essential for effective statistical analysis. In contrast, with a defective strategy or a poorly formulated question we can, at least, produce a conclusion, if not the correct one. The implication is that, in the past, understanding the mathematical and algorithmic detail was vital, if statisticians were to have any credibility, and had to take precedence over other issues. But, of course, this is no longer true. Nowadays, the necessity for understanding the details of the mathematics and algorithms has faded, at least for most users of statistics. Computers are taking over increasingly more of this role. This frees us to focus on more challenging and important strategic issues, such as question formulation. The importance of the role of mathematics in statistics in the past presumably explains why some people are unwilling to acknowledge that statistics is more than merely a branch of mathematics.

This has implications for the way that statistics is taught. Most teaching of practical statistics still starts at the level of having already decided what the question is. It then goes on to consider how to answer that question, focusing to a large extent on the mathematical manipulations—manipulations which will, in fact, be carried out by a computer. But this has skipped the most difficult and important stage, that of question formulation. Several institutions I know set exercises requiring the students to comment on particular published papers (from medical journals, for example). This is clearly a step in the right direction, though it is more akin to the apprenticeship notion (in which the statistician has to work in a particular environment to become expert in that area) than to formal teaching. More courses in statistical consultancy are needed. And these should begin at the beginning—before the statistical question has been formulated.

Finally, given the difficulty and importance of the problems of formulating statistical questions from the scientific questions presented, I would like to see more research effort spent on these issues: on the deconstruction of presenting questions so that the question being answered is that which was asked.

ACKNOWLEDGEMENTS

I would like to express my appreciation to Sir David Cox, Clifford Lunneborg, Shelley Channon, Kevin McConway, John Gower and the referees for their valuable comments on various drafts of this paper.

REFERENCES

Altham, P. M. E. (1984) Improving the precision of estimation by fitting a model. *J. R. Statist. Soc.* B, **46**, 118–119.

Brown, G. W. and Harris, T. (1978) Social origins of depression: a reply. *Psychol. Med.*, **8**, 577–588.

Carpenter, R. G. and Emery, J. L. (1977) Final results of infants at risk of sudden death. *Nature*, **268**, 724–725.

Chatfield, C. (1988) *Problem Solving: a Statistician's Guide.* London: Chapman and Hall.

——(1991) Avoiding statistical pitfalls. *Statist. Sci.*, **6**, 240–268.

Cox, D. R. (1977) The teaching of the strategy of statistics. *Bull. Int. Statist. Inst.*, **47**, book 1, 553–558.

Cox, D. R. and Snell, E. J. (1981) *Applied Statistics: Principles and Examples.* London: Chapman and Hall.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis (with discussion). *Appl. Statist.*, **43**, 49–93.

Early, D. F. and Nicholas, M. (1977) Dissolution of the mental hospital: fifteen years on. *Br. J. Psych.*, **130**, 117–122.

Efron, B. (1992) Jackknife-after-bootstrap standard errors and influence functions (with discussion). *J. R. Statist. Soc.* B, **54**, 83–127.

Everitt, B. S. and Smith, A. M. R. (1979) Interactions in contingency tables: a brief discussion of alternative definitions. *Psychol. Med.*, **9**, 581–583.

Finney, D. J. (1948) Main effects and interactions. *J. Am. Statist. Ass.*, **43**, 566–571.

Hand, D. J. (1986) Patterns in statistical strategy. In *Artificial Intelligence and Statistics* (ed. W. A. Gale), pp. 355–387. Reading: Addison-Wesley.

——(1990a) Emergent themes in statistical expert systems research. In *Knowledge, Data, and Computer-assisted Decisions* (ed. M. Schader and W. Gaul), pp. 279–288. Berlin: Springer.

——(1990b) Discussion of the papers by Edwards, and Wermuth and Lauritzen. *J. R. Statist. Soc.* B, **52**, 51–52.

——(1992) On comparing two treatments. *Am. Statistn*, **46**, 190–192.

——(1993a) Measurement and statistics: a Twentieth Century controversy. To be published.

——(1993b) Measurement scales as metadata. In *Artificial Intelligence Frontiers in Statistics* (ed. D. J. Hand), pp. 54–64. London: Chapman and Hall.

Haunsperger, D. B. and Saari, D. G. (1991) The lack of consistency for statistical decision problems. *Am. Statistn*, **45**, 252–255.

Herr, D. G. (1986) On the history of ANOVA in unbalanced factorial designs: the first 30 years. *Am. Statistn*, **40**, 265–270.

Holland, P. W. and Rubin, D. B. (1983) On Lord's paradox. In *Principals of Modern Psychological Measurement* (eds H. Wainer and S. Messick), pp. 3–35. Hillsdale: Erlbaum.

Kempthorne, O. (1980) The teaching of statistics: content *versus* form. *Am. Statistn*, **34**, 17–21.

Lord, F. M. (1967) A paradox in the interpretation of group comparisons. *Psychol. Bull.*, **68**, 304–305.

Lunneborg, C. E. (1992) A case for interpreting regression weights. *Mult. Lin. Regress. Viewpts*, **19**, 22–36.

Nelder, J. A. (1982) Linear models and non-orthogonal data. *Util. Math.* B, **21**, 141–151.

——(1992) Statistical packages and unbalanced data. *Comput. Statist. Data Anal.*, **14**, 403–406.

Oldford, R. W. and Peters, S. C. (1986) Implementation and study of statistical strategy. In *Artificial Intelligence and Statistics* (ed. W. A. Gale), pp. 335–353. Reading: Addison-Wesley.

Schwartz, D., Flamant, R. and Lellouch, J. (1980) *Clinical Trials.* London: Academic Press.

Searle, S. R. (1987) *Linear Models for Unbalanced Data*, p. 14. New York: Wiley.

Simpson, E. H. (1951) The interpretation of interaction in contingency tables. *J. R. Statist. Soc.* B, **13**, 238–241.

Stevens, S. S. (1946) On the theory of scales of measurement. *Science*, **103**, 677–680.

Tennant, C. and Bebbington, P. (1978) The social causation of depression: a critique of the work of Brown and his colleagues. *Psychol. Med.*, **8**, 565–575.

Wainer, H. (1991) Adjusting for differential base rates: Lord's paradox again. *Psychol. Bull.*, **109**, 147–151.

DISCUSSION OF THE PAPER BY HAND

J. A. Nelder (Imperial College of Science, Technology and Medicine, London): The topic of this paper is very important, because it is at the heart of the interaction between the statistician and the experimenter. The statistician must understand what the experimenter wants, and be prepared to find it misspecified in one of several ways. The opposite danger is that the experimenter will be presented with an analysis locked into a statistical framework that the statistician can cope with, rather than one relevant to the problem.

I hope that the speaker will bear with me if I do not use his terminology of deconstruction in my discussion. Deconstructionism is a barmy French literary theory of total relativism, whereas here relativism is decidedly *not* the name of the game. Also, deconstructing suggests to me only part of the activity required, namely pulling down a wrongly constructed edifice, without the second—and more important part—of constructing a better alternative. I prefer the term reformulation to describe both parts.

Among the interesting collection of examples there are three that raise general points of great interest to me. The first is example 6 (data in Table 4), a $2\times2\times2$ contingency table with one response factor and two explanatory factors. Of the two analyses quoted I think that the one using an additive model is simply wrong; additive models can give rise to negative fitted values, and hence do not satisfy the condition that models should not give impossible fitted values. The additive model also requires an interaction term, which the multiplicative model does not, and thus it fails the parsimony test. Here I assert that one model is wrong and one is right. What we are talking about here is the analysis phase, where the statistician's skills are most valuable. However, beyond that lies what I call the prediction phase, where we use the fitted values from a good model, together with their information matrix, to form quantities of interest and measures of their uncertainty. It is in formulating the quantities of interest that the aims of the experimenter become paramount, though the statistician may still have a part to play in clarifying the questions involved. The processes of analysis and prediction are to a large extent independent, save only that analysis comes first and prediction second. The literature is full of instances where the quantity of interest is formed first and then analysed (the use of signal-to-noise measures in quality improvement experiments is a good example: see Nair (1992)); this nearly always leads to trouble.

Example 5 (Simpson's paradox) illustrates the same point. The paradoxical inequalities are not severe here, but when they are this is always a sign that an interaction term is required in the model at the analysis stage, i.e. that the margins are not an adequate summary of the interior of the table. The question of weighting belongs to the prediction phase, and to the quantities of interest. We may want to give equal weights, or weights based on census data etc. The choice has nothing to do with finding a parsimonious model for the data at the analysis stage, but everything to do with answering the experimenter's questions (see Lane and Nelder (1982)).

Example 8 is rather different from the rest, for here it is statisticians who have built the edifice that needs to be torn down and rebuilt. Every day hundreds, maybe thousands, of experimenters are faced with the sort of output in Table 6, which is alleged to help them to make inferences from linear models.

For severely non-orthogonal data I assert that this output is almost useless. Type III and type IV sums of squares are based on confusions which I have described elsewhere (Nelder, 1977, 1982, 1993); type II sums of squares, A eliminating B and B eliminating A, can both be small when both A and B have effects, and a single type I analysis is quite inadequate to discover what is going on. Reformulation is urgently required, because I believe that the existence of this sort of output constitutes a statistical scandal. We should put our own house in order.

Finally I should like to give a brief reminiscence (relevant I hope). My introduction to medical statistics came via a doctor who posed the question 'how do I fit a skew distribution to my data, truncated at

TABLE 6

```
                    AMONG SUBJECTS (WHOLE PLOT) IS UNBALANCED
                        GENERAL LINEAR MODELS PROCEDURE
DEPENDENT VARIABLE: RESPONSE
                                              MEAN
SOURCE         DF    SUM OF SQUARES        SQUARE   F VALUE     PR>F    R-SQUARE       C.V.

MODEL          22     1107.68560606    50.3493473     22.01    0.0001    0.958432    15.4769
ERROR          21       48.04166667    2.28769841            ROOT MSE               RESPONSE
CORRECTED      43     1155.72727273                        1.51251394                 MEAN
 TOTAL                                                                            9.77272727

SOURCE         DF      TYPE 1 SS    F VALUE     PR>F    DF    TYPE 11 SS    F VALUE     PR>F

A               1    24.81893939      10.85   0.0035     1   30.00000000      13.11   0.0016
B               1    27.07500000      11.84   0.0025     1   27.07500000      11.84   0.0025
A*B             1    45.37500000      19.83   0.0002     1   45.37500000      19.83   0.0002
SUBJ(A*B)       7    56.45833333       3.53   0.0116     7   56.45833333       3.53   0.0116
C               3   921.54545455     134.28   0.0001     3  921.54545455     134.28   0.0001
A*C             3     6.66287879       0.97   0.4251     3    8.09629630       1.18   0.3413
B*C             3    15.92129630       2.32   0.1046     3   15.92129630       2.32   0.1046
A*B*C           3     9.82870370       1.43   0.2616     3    9.82870370       1.43   0.2616

SOURCE         DF    TYPE 111 SS    F VALUE     PR>F    DF    TYPE IV SS    F VALUE     PR>F

A               1    23.81332599      10.41   0.0041     1   35.68870656      15.60   0.0007
B               1    21.22477974       9.28   0.0062     2   34.18935006      14.94   0.0009
A*B             1    45.37500000      19.83   0.0002     1   45.37500000      19.83   0.0002
SUBJ(A*B)       7    56.45833333       3.53   0.0116     7   56.45833333       3.53   0.0116
C               3   907.12500000     132.17   0.0001     3  907.12500000     132.17   0.0001
A*C             3     6.34722222       0.92   0.4460     3    6.34722222       0.92   0.4460
B*C             3    13.27314815       1.93   0.1550     3   13.27314815       1.93   0.1550
A*B*C           3     9.82870370       1.43   0.2616     3    9.82870370       1.43   0.2616

TESTS OF HYPOTHESES USING THE TYPE 1 MS FOR SUBJ(A*B) AS AN ERROR TERM
SOURCE         DF      TYPE 1 SS    F VALUE     PR>F

A               1    24.81893939       3.08   0.1228
B               1    27.07500000       3.36   0.1096
A*B             1    45.37500000       5.63   0.0495
```

both ends?'. Me: why is it truncated? Doctor: because those people are abnormal. Me (distrusting the idea of abnormality): put them back and try taking logarithms. The distribution then looked symmetrical, so we tried a normal plot. It was straight right out to both ends (so much for abnormality). Summary: $\log y$ is normally distributed with mean m and variance s^2. However, beware! Not all experimenters like simplicity, believing, I suppose, that complexity is more respectable. What this says about their scientific training I am not sure.

I am most grateful to the speaker for delving into reformulation in so many areas, and for his analysis of the processes involved. I have great pleasure in proposing the vote of thanks.

Tony Greenfield (Little Hucklow): David Hand has identified the important problem of translating the research question into the statistical question. The first part of this problem is to discover what the researcher really wants to know. He described this as the tough part of the problem. Extreme care, he tells us, is needed in formulating research questions.

Extreme care is not enough. There are differences in language, in knowledge, and in perception of a subject being studied, between the scientific researcher who is the client and the consultant statistician. Even more important, but often not appreciated, there are differences in understanding of each other's interests and abilities.

My own consulting experience has presented difficulties of several types.

No research hypothesis

The researcher proceeds to collect as much data as possible and asks 'Let's see whether it tells us anything'. His perception of the statistician is that he can make sense of any data provided that there is plenty of it. An example was provided by a hospital registrar who, over several years, collected about 100 items of information on each of about 10000 kidney patients just with the hope that it might prove useful.

Many hypotheses implied

A group of researchers aim to answer all possible questions in one great all-embracing trial. An example was a multicentre study of cot deaths in which about 1800 variables were recorded on each of about 1000 deaths and about 1000 controls. The study was not designed according to published research hypotheses. Hypotheses were not stated except implicitly by the questions on the data collection forms which were designed by groups of people sitting round a table and plucking ideas out of the air. The researchers assumed that each question could be analysed separately by cross-tabulation and associations tested by 'standard statistical methods'. This study was a great waste of public funds.

Statistician is client's servant

In some cases the researcher believes that he knows enough about statistics to be able to prescribe the appropriate analysis. He tells you that he wants a t- (*or* χ^2-) test, or a straight line fitting. He may even think about the statistics before he starts his trial and ask for a power calculation. 'But I'm asking for only a few minutes of your time', he will plead. Then, when you question his approach and start probing deeply, he will wonder whether he has asked the right person. 'You are a statistician?', he might ask.

All these situations support the oft-repeated entreaty: 'Consult a statistician before you design your study'. If you can persuade your client to do this you are on track but you will lose track by launching straight into the question 'What is the problem?' because the way that he answers this will depend on *his* perception of *your* subject. He may feel that he cannot describe some aspects of his research because they would be beyond your understanding.

The solution is to be honestly ignorant, openly naïve and eager to learn from your client, to be able to see the situation from his viewpoint. You need to ask him what it is about and encourage him to explain the underlying chemistry, engineering or physiology. You should ask what knowledge exists and what research hypotheses have already been formulated and tested by experiment.

A suggestion that I have fruitfully made on several occasions is for the client to draft his final report before he even drafts the protocol of the research to be done, encouraging him to include tables and graphs of data that he expects the study to generate. Only then might he appreciate the importance of writing a protocol, including a statistical protocol, for his research.

Douglas Altman, in a recent personal letter, wrote:

> 'In my experience, protocols are rarely produced in medical research except when designing clinical trials or when applying for a grant. In other words, most medical research is done without protocols. This has clearly been the case when one has had to spend half an hour or more trying to extract from a researcher what he thought his study was all about.'

The same may be said for any other field of study. But if, through education, example, and discussion, using the methods I propose, the clients can be persuaded to write detailed protocols, then we shall have begun to tackle the problem posed by David Hand.

Another suggestion which I sometimes offer to clients is to simulate a study before it is done. This is easy with an experimental design program, such as DEX, which can be used to record a researcher's expectations and to demonstrate the range of models that might be fitted to his data.

Another problem relates to the researcher's perception of the statistician. This is his fear that the statistician will make difficulties by being too complicated. He will ask you: 'Please keep it simple. I just want to test one variable at a time'. This attitude is difficult to overcome but it must be faced when a multivariate analysis is clearly needed. A few simple examples ready to hand will help. The following usually impresses.

In the blood there is a clotting factor VIII-C and Fig. 1 shows the values for two sets of people. The lower line shows values for some women who are known to be carriers of haemophilia and the upper shows values for some women who are known to be normal. Although the averages of the two sets of values may be different, there is much overlap.

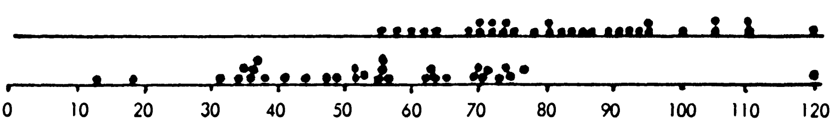

Fig. 1. Dot diagrams for clotting factor VIII-C values for two sets of people: haemophilia carriers and normals

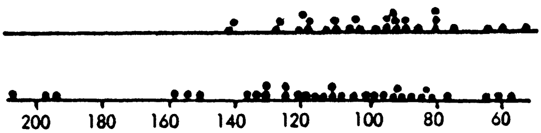

Fig. 2. Dot diagrams for clotting factor VIII-RA values for two sets of people: haemophilia carriers and normals

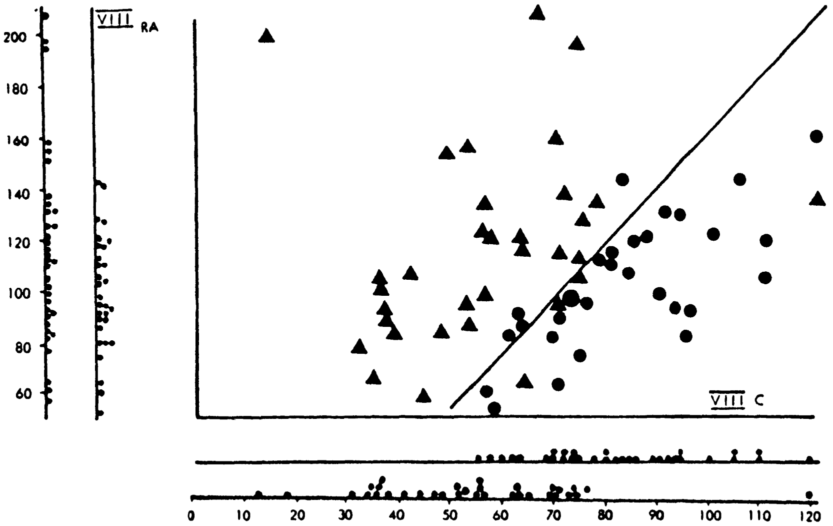

Fig. 3. Bivariate plot of the values shown in Figs 1 and 2, with a discriminant line almost completely separating the two sets of values: ▲, carriers; ●, non-carriers

The fact that a statistical test may show that the averages are different is no consolation whatever to the poor woman who wants to know whether or not she is a carrier and her value of factor VIII-C is somewhere in the middle of the total spread.

Fig. 2 shows the values of factor VIII-RA for the same two sets of people. There is even more overlap than with factor VIII-C: so much that the clinician might argue that there is no value whatever in considering this extra measurement. However, if we plot the two values together on a single graph, factor VIII-RA against factor VIII-C, a new picture emerges as in Fig. 3. The overlap has almost disappeared and a good diagnostic rule can be found by using two variables that tested individually might have been discarded as useless.

David Hand's paper has opened an important debate and I have pleasure in seconding the vote of thanks.

The vote of thanks was passed by acclamation.

Hans J. Lenz (Free University of Berlin). Presenting a few examples and theorems we vote for sound models, clear hypotheses and carefully checked data. This view was originated by Morgenstern (1963). Its revival is due to David Hand.

Firstly, consider the theory of *linear aggregated consumption* (see Prais and Houthakker (1955)). Let y_s be the consumption, x_s the personal income of household $s=1, 2, \ldots, N$ and the consumption model be given by $y_s = a + b \log x_s$. Aitchinson and Brown (1957) assumed $\log x_s \sim N(\mu, \sigma^2)$ for all

$s=1, 2, . . ., N$. Let $Y=\Sigma_{s=1}^{N} y_s$. Then $Y=Na+bN\log\mu^*$ where $\mu^*={}^{N}\sqrt{\Pi_{s=1}^{N} x_s}$. However, the geometric mean $\mu^*=\exp\{2\log\mu-\log(\mu^2+\sigma^2)/2\}$ is linked in an unpleasant way to the parameters of $N(\mu, \sigma^2)$ because model specification and aggregation are not compatible.

It is interesting to ask when a linear aggregation is perfect, i.e. when a homomorphism exists given a model $(\mathbf{A}; \mathbf{y}, \mathbf{x})$, a transformation T: $\mathscr{R}^H\rightarrow\mathscr{R}^J$ and a statistic S: $\mathscr{R}^F\rightarrow\mathscr{R}^G$; see Fisher (1962), Schneeweiss (1965) and Sondermann (1973). Dropping any latent errors in the equations we have $\mathbf{y}=\mathbf{Ax}$ as a micro model, $\mathbf{X}=\mathbf{Tx}$ as a linear aggregation, $\mathbf{Y}=\mathbf{BX}$ as a macro model and $\hat{\mathbf{y}}=\mathbf{SY}$ as the predictor. Then $\hat{\mathbf{y}}=\mathbf{y}\Leftrightarrow\hat{\mathbf{y}}=\mathbf{SBT}$ *or* $\mathbf{B}=\mathbf{S}^+\mathbf{AT}^+$ where $\mathbf{S}^+$ and $\mathbf{T}^+$ are Moore–Penrose inverses of $\mathbf{S}$ and $\mathbf{T}$.

Secondly, *marginalization* in multiway tables must be carefully used. Let $\mathbf{f}$ represent the frequency of professors in the individual faculties of a given university. Assume $\mathbf{f}=(f_1, f_2)=(20, 20)$. Note that the total number is 39. A theorem exists that says that f cannot be summarized if the underlying relationship faculty ↔ professor is m:n. I believe that there are many artificial intelligence specialists who propagate a fuzzy set approach to the data (20, 20, 39)!

Thirdly, counter-intuitive effects can be generated by *temporal aggregation*. Consider a microtime $T=\{0, \pm 1, \pm 2, . . .\}\in\mathbf{Z}$ with granularity 1 and a macrotime $T^*=\{0, \pm\delta t, \pm 2\delta t, . . .\}$ with granularity $\delta t\in\mathscr{N}$. Let $\delta t\rightarrow\infty$. Then $\bar{H}\rightarrow\infty$ where $\bar{H}$ is an upper bound of the maximum lag $H\in\mathbf{Z}$ in a distributed lag model (Schönfeld, 1979; Werner, 1982).

Tiao (1972) has proven that the first differences of the monthly averages follow an MA(1) process independently of the number m of subperiods used for temporal aggregation if the generating process is a simple random walk. Kirchgässner and Wolters (1992) confirmed this empirically for $m=2$. Christiano and Eichenbaum (1987) have shown that, when averaging a continuous ARMA(p, q) process with $q<p$, the macrotime generation process is ARMA(p, p).

Finally, Lütkepohl (1987) proved further theorems on the reproduction of vector autoregressive processes if specific vector stochastic processes are linearly aggregated.

Chris Chatfield (University of Bath): I welcome this paper which tackles a topic of special interest to me. Most of the statistical literature is directed towards *techniques* rather than *strategy* but it is the latter which is much more difficult to learn and to give general guidance on. For example, it is easy to *do* regression, but much more difficult to know *when* regression is appropriate and *which* model to use.

But although I liked the paper, I have to say that I do not like the title. *Deconstruct* is not a word which appears in my dictionary (although I suppose that, like Humpty Dumpty, we can define a word to mean anything we want it to, and it does see controversial use in the study of English language). The trouble is that the word sounds rather negative. I agree that we sometimes need to strip a problem down to its component parts, but we then need to put it back together, and I think that problem formulation should primarily be a positive, constructive exercise. Thus I would prefer to talk about 'formulating statistical questions' as the author himself does in Sections 5 and 6.

The paper gives some thought-provoking examples (Section 4) and a useful start at establishing general principles (Section 5). However, the process can never be made fully objective and will continue to depend largely on experience.

The main points that I would add are as follows.

(a) Be prepared to ask many probing questions to ensure that sufficient background knowledge is available and to make sure that the correct objectives have been specified (see Chatfield (1988), chapters 3 and 10).

(b) Always ask to see any data that have been collected. This is especially important in dealing with the 'scientist who knows what he wants' (see example 1 in Chatfield (1991)).

(c) Avoid answering questions over the telephone, as I have found this a recipe for disaster. In particular you cannot see the data when talking on the telephone.

What are the implications of this paper for *teaching*? Students are usually required to answer prespecified, unambiguous clear questions, which, as noted by the author, skips the most difficult stage of real life problems. I have sometimes tried setting questions to more advanced students which are deliberately incomplete or even partially wrong. Although this is rather unkind in some ways, the students learn fast that they must ask questions and should not take things for granted!

Donald A. Preece (University of Kent, Canterbury): Professor Hand speaks of the questions that the researcher wishes to consider. These are often three in number:

(a) how do I obtain a statistically significant result?;
(b) how do I get my paper published?;
(c) when will I be promoted?

So Professor Hand's suggestions must be supplemented by a recognition of the corruptibility and corruption of the scientific research process. Nor can we overlook the constraints imposed by inevitable limitation of resources. Needing further financial support, many researchers ask merely 'How do I get results?', meaning by 'results', not answers to questions, but things that are publishable in glossy reports.

Where do statisticians come into this?: perhaps nowhere. Quite recently, a research council decided that its statisticians were 'service staff'—which indeed they were, as servants of colleagues and of science—but 'service' as opposed to 'scientific'. Anybody who knew which way up to hold a test-tube was a scientist; but a statistician—that incomprehensible mathematician and computer wallah—was not, and so could offer nothing to research planning. Alas, Professor Hand lends credence to this with his false dichotomies between 'researchers and statisticians' (Section 1, eighth paragraph) and between 'scientific and statistical questions' (Section 8, first paragraph). This tends to confirm the statistician as a mere outside consultant whom people perhaps cannot afford until they are in a mess, by which time a statistician is needed to paint respectability over defective work. No, as Box (1993) stated, 'the statistician must strive to earn the title of first class scientist'. Nowadays, a statistician who did this might be considered by other scientists to be too uppity to be tolerable as a colleague, but we must work on it. . . .

In practice, who is 'the client'? Usually, the client is not an individual, but a team, committee or board, whose different members have different aims and questions. Any statistician involved should be expert in the meaning and achieving of compromise, and in recognizing which aims will not be met, which questions not answered, by any particular compromise.

Professor Hand sometimes implies that a client has just one question, notwithstanding R. A. Fisher's famous suggestion that 'Nature . . . will best respond to a logical and carefully thought out questionnaire' (Fisher (1926), discussed by Preece (1990)). Questions range from the general ('What is going on here?') to the specific. I would have been happier had the paper included linguistic and philosophical scrutiny of the concept of a 'question'.

Clifford E. Lunneborg (The Open University, Milton Keynes): Professor Hand is to be commended for directing our attention to the relative ease with which statisticians' answers may fail to engage researchers' questions—or, the questions researchers ought to ask. I offer the following example to suggest that we also can be led astray when the researcher has done something so intelligent that we forget to ask 'Why did you do that?'.

Pursuing my interest in bootstrap inference I found in Sprent's revised nonparametric text (Sprent, 1993) an excellent introduction to the bootstrap, supported by this example. A researcher confronted with the data 0 6 7 8 9 11 13 15 19 40 computes a 10% trimmed mean as a location estimator, removing the 0 and 40 before averaging. Sprent illustrates drawing bootstrap resamples from these data and studies the distribution of 10% trimmed means computed from 100 such resamples. He reports considerable skewness in that distribution and a variance not much reduced from the bootstrap variance for the untrimmed mean and concludes that, for the example, 'bootstrap estimation undoes some of the robustness that one attempts to build in by trimming'.

What went wrong? Sprent notes that the asymmetry and high variance of the bootstrap distribution is due to the fact that in a certain proportion of the resamples two or more 40s will be present, only one of which will be trimmed before computing a mean. Better bootstrap estimates (of bias and standard error) would result, he notes, if 20% trimmed means had been computed. The researcher, though, apparently chose to compute a 10% trimmed mean.

An important insight to the choice of estimator and, I believe, to how inference might better be pursued is given by Sprent's comment: 'inspection of the original sample suggests that the prime reason for using the trimmed mean is to downweigh the observation 40'. If this is true—and asking 'Why?' would be desirable here—it would mean that the researcher was not committed to computing a 10% trimmed mean. That estimator was chosen precisely because there was a single 40. Had there been no 40 (or similarly 'large' value) an untrimmed mean would have been computed and had there been two such large values a 20% trimmed mean might have been used. Were the researcher known to be following such an adaptive strategy in computing an estimate, then any inference drawn about the accuracy of the estimate should take that strategy into account. If the inference is resampling based then the strategy

TABLE 7
Fuel efficiencies of two types of car

Researcher	*Efficiencies for type 1*			*Efficiencies for type 2*		
	Car 1	*Car 2*	*Geometric mean*	*Car 1*	*Car 2*	*Geometric mean*
English (miles per gallon)	1	4	2	2	2	2
French (gallons per mile)	1	0.25	0.5	0.5	0.5	0.5

should be applied to the resamples and not simply the calculation of the particular statistic derived from the original sample, no matter how appropriate that statistic for the original sample.

M. C. Jones (The Open University, Milton Keynes): On example 3, first, the example about miles per gallon and gallons per mile, what is needed is an 'average' that commutes with the reciprocal transform. The geometric mean will do the trick (or equivalently averaging the logarithms). Table 7 is a new version of Table 2 based on the geometric mean. Here the English and French 'averages' are the reciprocals of one another as desired. And so, in the interests of European unity, both French and English come to the same conclusion. (And that conclusion happens to be that the two car types are the same.)

Here is a closely related point. There is a passage in Section 4.1 of the paper which talks about measurement scale influencing choice of method. In relation to ordinal scales, it says that nonparametric test people switch from means to medians. Certain words, 'unfortunately' and 'at the cost', perhaps inadvertently, make this sound like a cheat when, as David said, it is surely the right thing to do, since medians are invariant to monotone transformation, means are not. I say, good for the nonparametric testers: do not worry about changing from the badly formulated hypothesis that is the one concerning means. The problem in all this contribution so far, of course, is that people jump to arithmetic means—which are tied up with linearity, additivity, normality, etc.—when they really mean 'location' or 'typicality' in some general sense.

On example 6, next, the example to do with comparing probabilities, Professor Nelder has already explained why looking at differences is just not right. Ratios are better, but I am worried about multiplying probabilities too. Take 1 minus all the probabilities in the table and the ratios become 0.71 and 0.90 (which may or may not give different indications). The problem is tied up with looking at, say, the bottom row of Table 5 (thus conditioning on 'intimacy' people only), and claiming that depression is 11 times more likely for 'event' people than for 'no event'. Take 1 minus things. Then depression with 'no event' is only '1.11 times as unlikely' as with 'event'. As my colleague Fergus Daly made me aware some time ago, odds are fine in this regard. The two odds ratios turn out to be 13.8 and 12.2, and, if we considered proportions without depression instead, we would have their reciprocals.

Finally, I thank David for a very interesting and enlightening paper, and I say that not just because his office is next-door-but-one to mine!

John Gower (Wheathampstead): It is high time that the Society discussed some of the issues underlying statistical consultation and I therefore welcome this paper. The precise meaning of *deconstruction* in the statistical context is elusive; it seems to mean that one should look closely at the assumptions of model, design and hypotheses that underlie statistical analyses and view these in the light of what questions, if such are discernible, the researcher is asking. These are, of course, fundamental issues but the only deconstruction *process* that I see is for the consultant to ask some questions of his own. One of a consultant's most useful contributions is to know what questions to ask and Professor Hand's paper helps here, though I do not discern any really general principles in Section 6; more, useful pointers.

An area I would highlight is the choice of model. Only the scientist himself can specify, perhaps with some encouragement, substantive models; most competent scientists can specify several plausible models. On statistical grounds alone it is very difficult to distinguish between models with the amount of data collected or which can be afforded. Although fascinating to statisticians, random variables which merely model noise or deficiencies of measurement are tiresome to researchers, who often misunderstand the role of statistical descriptive models and may dismiss those that obviously do not model any familiar scientific process.

The crucial problem is the interaction between scientist and statistician. Each has much to learn from the other. A statistician cannot advise properly unless he knows something (the more the better) of the substantive science and, for this reason, fruitful statistical collaboration should be a continuing process over many years (Gower and Payne (1987) report an example of the benefits of long-standing collaboration). Whether or not this implies that statisticians should be educated within a specified application area, as happens in the USA, for example with medical statisticians and psychometricians, is an interesting question. It produces statisticians with a good understanding of their application areas but it tends to be divisive in producing parallel terminologies and duplication of methodological developments which can be a source of confusion to others. If statistics is to be regarded as a discipline, one of its strengths is its set of common principles, a thorough understanding of which is required to sort out the examples listed in Section 3 of Professor Hand's paper and which inform the asking of appropriate questions.

Richard A. Stone (City University, London): I commend this investigation of how the formulation of a research question relates to the application of statistical techniques.

There are implications for the nature of the interaction between statisticians and scientists. Let us view the most common situation, the scientist as a client committing valuable resources to consult an expert statistician, from the scientist's perspective. If the statistician is good he will question the basis of the research objectives and risk being seen as a pompous 'keeper of the true scientific method'. How many scientists are game to come back for more of this?

The solution is for statisticians to be colleagues of scientists, performing defined roles in research teams. After all, much of scientific or engineering research is already interdisciplinary. The other team members should know enough statistics to understand the methods that are being applied, but need not be statistically creative. This is an ideal which would be shared by many, but there is another requirement that is rarely considered.

This is for the *managers* of research, the people to whom teams report, to have sufficient statistical training to enable them to understand and be critical of the application of statistical methods *and* to give them an appreciation of the metastatistics in asking the right research questions. Without this statistical pull from managers, there is usually little incentive for a research team leader to include a statistician.

In summary, if it is important for practising scientists and engineers to have a good training in statistics, it is that much more important for managers to be statistically literate and able to question deeply in the spirit of this paper.

The author's deconstruction of the comparison of two treatments A and B sits within a broader framework. A measure of the difference between the two potential responses for a subject (such as $z_A - z_B$) may be viewed as the most fundamental sort of causal effect. Statisticians are generally comfortable with the notion that randomized experiments allow causation to be tested or estimated, but their analyses almost always concern the expectation of unit-level causal effects (such as $\mathscr{E}(Z_A - Z_B)$) without recognizing the loss of generality.

The most general sort of causal effect which can be tested or estimated is some measure of the difference between the distribution of Z_A and the distribution of Z_B, which may be termed a *distribution* causal effect (Stone, 1993). Hand notes that the distribution causal effect $\Pr(X_A > Y_B)$ which underlies the Wilcoxon test is not the same as the 'desired' quantity $\Pr(Z_A > Z_B)$, but there is an analogous limitation with expected differences—the measure $\mathscr{E}(Z_A - Z_B)$ can be 0 even though unit-level causal effects exist.

M. C. Fessey (Newport): I do not know what sampling fraction justifies Professor Hand's opening assertions that *much* statistical analysis is misdirected and that *many* statisticians pursue mathematically misleading solutions to problems. But I have an intuitive sympathy with his view that developing strategies for asking the right questions should have precedence over inventing mathematical tools to solve them.

Do Professor Hand's strictures apply also to what macroeconomic statisticians do? Do many of them often ask the wrong question or formulate the right question wrongly or apply inappropriate tools to answer it?

Even more than the scientific examples Professor Hand adduces, economists' questions are a reflection of fashion rather than derivatives of the application of the kind of strategies that he outlines: Keynes in the seventies; Friedmann in the eighties.

In economics, errors in questions or the absence of strategies or the choice of inappropriate tools may have less effect than factors which play no part in Professor Hand's exposition. For example economists do not design experiments badly; their questions are about the world as it exists.

And what contributes to the poor reputation that economic scribblers enjoy is that the information the world provides is rarely the information that strategies require. Value added is the true measure of national activity. But in most cases we have yet to discover how to measure value added. So, instead, measure production. But often we cannot measure production. Instead, deliveries?: but these are often not available. We make do with sales.

Then too a guess at value added must pay regard to stocks; I am not sure that we have ever considered from first principles how to measure the value-added component of stocks. But that aside, we have to make what sense we can of what we know of the conventions of first in–first out and last in–first out and so on. And after all that we have estimates of value added in producing goods and services at current prices—terms which trip off the tongue but whose meaning varies. No wonder that after Moser's rule that any interesting figure is probably a mistake comes a second precept: no two sets of figures ever agree.

But, to return to deconstructing economic questions, the last time thought was given to what questions are relevant to the broad field of economic analysis was 30 or so years ago when financial statistics burgeoned at the turn of the 1960s in response in this country to the Radcliffe report on the working of the monetary system. Perhaps the time is ripe for a search for a strategy for linking macroeconomic questions as well as scientific questions to the statistical questions designed to answer them.

S. J. W. Evans (London Hospital Medical College): I want to thank Professor Hand for a paper which is as excellent as its title is terrible! I have been teaching a very short course in statistical consulting for the Master of Science in medical statistics at the London School of Hygiene and Tropical Medicine for a few years, following my experience of running a statistical 'clinic' for doctors and medical researchers at the Royal London Hospital for 15 years.

A non-statistical aspect of such consulting is the skill of listening to the 'patients' (i.e. the doctors—we reverse roles in my clinics!) and asking the right questions. At the design stage these concentrate on what they really want to find out: at the analysis stage on *exactly* what they have done. I have found the set of questions by Mainland (1964) to be a useful framework.

A major problem is that the types of personality for whom mathematical statistics has appeal are frequently those for whom communication skills have little appeal.

Rather than requiring that all statisticians are capable of exercising all skills it seems that what happens, both in practice and what should happen, is that there must be a spectrum of statisticians from the very mathematical to those for whom answering scientific questions is of the greatest importance. They will need training in science and in communication skills. As noted elsewhere (Evans, 1991) medical statisticians in our department are encouraged to visit laboratories, to look down microscopes, to visit out-patient clinics and to attend ward rounds in hospital.

In addition they will have to be taught some of the skills of a counsellor in being able to put a client at ease. These questions are not always welcomed especially when the investigator is fearful or wishes to impose a merely technical role on a statistician.

There are a variety of issues in deciding what is the best method of analysis to answer a particular question, and frequently the vital component is to decide on what scale it is that the question has greatest clinical or practical relevance. Some of Professor Hand's problems will be solved by Martin Bland's adage, 'take logarithms and do a *t*-test', which is said to be what all statistical consultancy is about!

Toby Lewis (University of East Anglia, Norwich): Professor Hand's important paper raises fundamental issues which have resonances outside the various contexts treated in the paper. I suggest that his thesis applies directly to a vitally important field, that of *assessment*—examining, testing and performance evaluation in its many manifestations. This is essentially a statistical activity, because it is *estimating* something—skill, knowledge, understanding, performance.

Professor Hand raises the central issues 'What question do people really want to ask?' and 'what question *should* they want to ask?'. He says in his paper (Section 1; my italics):

> 'The aim of this paper is to stimulate debate about the need to formulate *research* questions sufficiently precisely that they may be unambiguously and correctly matched with *statistical* techniques'.

If we substitute for the italicized words 'research' and 'statistical', Hand's remarks apply directly to the need to formulate *performance evaluation* questions with precision and to match them unambiguously and correctly with *assessment* techniques.

In Section 2, he writes (my italics):

> '. . . I am concerned with identifying . . . what it is the *researcher* wants to know—an aspect of *statistics* which precedes the choice and application of techniques . . .'.

Again, substituting for the general words 'researcher' and 'statistics' we have a statement of the importance of identifying what the *performance evaluator* wants to know, an aspect of *assessment* which precedes, or rather ought to precede, the choice of techniques.

All over the world, all the time, millions of people are being examined and assessed and are taking tests, while others are preparing and administering those tests. The cost and effort are vast—but is all of this assessment essential? Has it been thought out? Where needed, is it being done in the most effective way? To what extent has the activity become just a habit?

Hand's far reaching thesis is of direct application to these issues. May one hope that he will give us his thoughts on this, perhaps in the form of another discussion paper before long?

The following contributions were received in writing after the meeting.

A. S. C. Ehrenberg (South Bank Business School, London): David Hand rightly asks us to think about our statistical problems. But if this is 'deconstruction', then I have been talking prose most of my life.

More seriously, I cannot agree with Professor Hand's view that the choice of what question to answer 'depends on what the investigator wishes to know'. Investigators cannot decide to ignore potentially major factors because 'they do not wish to know'.

To illustrate with his first example in Section 3.1, this was about a two-group clinical trial of

(a) a radiotherapy treatment *versus*
(b) the radiotherapy preceded for 30 days by a sensitizing drug.

Hand notes two possible experimental designs. In the *pragmatic* design the radiotherapy-only treatment starts when the 30-day sensitizing drug starts in the second group: this could not tell us how far any apparent effect is due to the drug as such or to the 30-day delay. In the *explanatory* design the radiotherapy-only treatment starts 30 days later, i.e. as for the group who also have the sensitizing drug: this would evaluate the sensitizing drug but we would learn nothing about the 30-day delay.

But is it merely up to the analyst or researcher to decide subjectively which design to use and hence what conclusions to draw? An alternative is to use *both* designs and to cover all the main factors in the situation. Or can we not deconstruct the problem a little more and run a *three-group* trial? Average subsample sizes need only be reduced by $\frac{1}{3}$ (and, since Student's *t*-test, classical statistics has let us deal with very small samples!).

A further worry is the apparent absence of prior knowledge. What sort of investigator is it who does not already know beforehand roughly what effect (if any) a 30-day delay in such a radiotherapy treatment would (or would not) tend to have? And surely no potentially successful treatment is evaluated and/or approved on the basis of just one clinical trial?

David J. Finney (Edinburgh): This paper should encourage all who teach statistics, or who are ever consulted on statistical matters, to think critically about the fundamentals of *being a statistician*. We have been taught, and have taught others, much about the mechanics of analytical techniques; rarely has there been sufficient emphasis on the strategy of choosing a technique appropriate to a particular problem. Computer software sometimes aggravates the evils to which Professor Hand draws attention by implying that its user need do no more than to feed data into his personal computer to receive answers to possibly inexactly specified questions and even to be provided with output in good shape for publication.

Even the present paper could seem to support the belief that the most important task for a statistician is to report results of tests of significance: this idea is welcome to editors as enabling authors to state conclusions in very few words, often without statement of mean values or parameter estimates! Medical journals today often convey the impression that in clinical research the only relevant question is 'Did the treatment have a significant effect?'. My own special interest in biological assay can illustrate the need for strategic balance. Typically, for two materials S and T, there is reason to adopt a model stating that an experimental dose z of T will behave exactly as does a dose ρz of S, irrespective of the magnitude of z or of the details of the experiment. The object of an experiment is to estimate ρ as precisely as possible; a significance test on the adequacy of the underlying model may be vitally important. Much is known about experimental designs and the choice of responses to be measured for achieving test power and precise estimation: pharmacologist and statistician must collaborate in balancing the two.

Professor Hand makes a valuable distinction between 'multiple univariate' and 'intrinsically multivariate'. This deserves further comment: many sets of data, whether from experiments or from surveys, are essentially multivariate in the information recorded yet an analysis that produces a principal component or other composite index may fail to illuminate important questions concerning one or more variates separately. If all users of statistics followed the practice of chemists in their care for terminology, ambiguities might be reduced; I observe an increasingly sloppy practice of regarding 'slope' as a synonym for 'linear regression coefficient', and even 'odds' for 'probability'.

A. M. Herzberg (Queen's University, Kingston): Professor Hand must be congratulated for his contribution to a discussion of an important topic. He gives it a name, i.e. deconstructing statistical questions. Will this word 'deconstructing' soon be in every statistician's vocabulary? Will statisticians become known to the general public as 'deconstructistics'? Will this make us better off at parties than Sir Claus Moser's unpopular statistician (Moser, 1980)?

More seriously, perhaps, Professor Hand's paper is related to the issue of statisticians learning about other disciplines. Do we not need more people like Sir Ronald Fisher and Sir Harold Jeffreys who made use of statistical inference in their work in the natural sciences and found the link a two-way street. Perhaps we also need to educate and train 'scientific generalists', a term coined by Bode *et al.* (1949). A summary of their paper might be considered to be

> 'Recapture the universalist spirit of the early natural philosophers.
> Learn science and not sciences.
> Know in capsule form the dozen central concepts of each of the major sciences.
> Learn in the habits of mind of the chemist, psychologist and geologist.
> Use in each science some of the intellectual equipment of the other sciences.
> Be exceptional in breadth of appreciation.
> Be able in biological and medical science to suggest physical explanations of mathematical models for known or conjectured facts.
> Be familiar with forging and milling, the functions of a turret lathe . . .'

(summary quoted from MacLay (1991)).

P. Lovie (University of Keele) and **A. D. Lovie** (University of Liverpool): Professor Hand's paper must make uncomfortable reading for almost everyone who has been called on to give statistical advice. He is right to remind us that research questions are paramount and it is these that must dictate the statistical strategy, and not vice versa. A deconstruction approach, such as is suggested, will go a long way towards remedying current wrongs in formulating scientific hypotheses, but *unambiguous* identification of the researcher's aims and intentions is likely to be a far trickier, if not impossible, task—for neither client, nor even statistician, ever comes 'naked and with nothing' into the negotiations which inevitably characterize relations between the two. Researchers invariably bring both implicit, or tacit, presuppositions and explicit expectations about the outcome of any investigation, or about what statistical analyses might be appropriate. The statistician may have certain preconceptions about the sort of problems likely to be encountered in the client's domain or preferences for particular statistical approaches. Each will have their own interests to satisfy, which do not necessarily wholly overlap.

Having argued this far, it is possible to suggest moving in an even more radical direction, although without interpreting deconstruction quite so nihilistically as Derrida (see, for example, Hoy (1985)). Ethnomethodologists, most notably Garfinkel (1967) (see also Gephart (1988)), have advocated the use of 'breaching' strategies which challenge the unreflected-on realities. The parallel here is that the statistician would insist that the client should explicitly justify the choice of variables, measurements, designs, outcomes and also give a proper accounting of the background theory including alternative and competing points of view. The trick is to develop challenging strategies which involve both sides in a creative and constructive negotiation.

A strongly related issue is the contribution that the statistician might make to the rhetoric of the client's case. In other words deconstruction leads to reconstruction where the statistician can assist in reassembling the outcomes of the negotiation into the most persuasive form achievable.

In this bold paper Professor Hand challenges us to rethink the statistician's role in scientific change; this debate must surely continue.

R. J. Mackay and R. W. Oldford (University of Waterloo): We agree that more effort needs to be spent on problem formulation and that this is an important part of statistics.

The impact this has on teaching is significant. Our first course now presents statistics as the methodology of empirical problem solving consisting of five broad steps—problem, plan, data, analysis and conclusion—each one of which is dependent on the previous steps. Each step has its own focus and can be broken down into a number of substeps which must be taken in any application.

The 'problem' step attempts to deal with many of the issues raised in the present paper. Some useful concepts and terms are the units and target population (or process), response and explanatory variates, and population attributes of interest. Many of the problems raised in the examples can be discussed entirely in these terms. Note that population attributes are defined, often conceptually, by considering the behaviour of the response over the target population. These can include means, percentiles and similar quantities conditioned on the values of explanatory variates. The fundamental questions of the researcher must be translated into questions about these attributes.

For example, Lord's paradox as presented in Section 3.4 can be described in these terms. The units in the target population are undergraduates, the response variate is a person's weight and two of the explanatory variates are the person's sex and the time at which they are weighed. If the attribute of interest is the average difference in change of weight between males and females, then the first analysis proposed is fine and will lead to the conclusion of no difference. If the attribute of interest is the average difference in change of weight between males and females of the same initial weight, then a better study would be to include only those males and females whose initial weights can be matched with that of a member of the opposite sex. Then the noticeable difference would be picked up by either analysis method because the initial weight means would be identical for both sexes.

In some cases, it may be found that the population attributes are fundamentally inestimable. An example is the population proportion of times $z_1 - z_2 > 0$ as discussed in Section 3.2. Conflicts between what is desired by the investigator and what is statistically possible must be resolved before proceeding to the collection of data.

There are several ramifications for teaching that follow from this approach. We need to provide students with a new vocabulary to discuss issues in each of the steps. A detailed description of context is required in all examples so that enough information is available to work through the process. Far less time is spent on the 'analysis' step than in a traditional course.

I. W. Molenaar (University of Groningen): It is a great pleasure to reply to Professor Hand's paper. I dislike his destructive neologism 'deconstructing', but I welcome his initiative to focus attention on the careful formulation of the right statistical question. Its importance is grossly underestimated, both by statisticians and by clients.

In the Netherlands the generation–testing distinction goes back to Hemelrijk's (1958) paper about 'statistical detection' and 'statistical proof'. I have expressed my concern in public about pure exploration (for data, though it have no tongue, will speak through most miraculous organ) and emphasized considering both the research goal and the available prior knowledge before deciding what data to collect and how to analyse them (Molenaar, 1988).

The first example of Section 3.1 has left me confused. I would say that there are three treatments: immediate radiotherapy, radiotherapy after waiting 30 days and radiotherapy after 30 days' drug use. It may be unethical to apply the second treatment, because a patient in need of radiotherapy will probably deteriorate during the waiting period. Measuring the patients' condition at day 0 and 30 would be helpful, and when a differential effect for different pretreatment scores is plausible one would consider random assignment to pairs or triples matched on this score. Moreover Hand's D2 design only helps to 'understand, or explain' if we have reason to believe that there will be a differential effect depending on some patient variables and/or some theory about the effects of the drug, and the relevant covariates are included in the study. If restriction to homogeneous patient groups is helpful, this may imply that future use of the best therapy for more heterogeneous patient groups is not supported by the conclusions of the study.

Example 2 relates to work by Holland (1986) and Rubin (1991) treating the score of the subject on the treatment that (s)he did not receive as a missing observation. Many researchers present both the t-test and the two-sample Wilcoxon test only for the pure location case (seldom found in practice) of a uniform shift. In a medical setting the conclusion 'with A more people improve than with B' could be enough to choose A, although fairly often the amount of improvement would also matter. Moreover, the study would be more valuable if statistical detection was added to look for subgroups who might be better off with B. In an agricultural or industrial setting the 'population' view of Section 5.1 would nearly always prevail, because the total gain is what matters.

At the end of Section 5.5 I was surprised to find only a multivariate analysis of variance, and not a combination of all indicators into a univariate scale.

Peter C. O'Brien (Mayo Clinic, Rochester): The need to tailor study design and data analysis to answer the scientific question is, I believe, profoundly important. I provide some further examples which occur frequently in medical research.

One-sided versus two-sided tests

A common view is that 'the test should be two sided if the investigator would be interested in a difference in either direction'. Since investigators are interested in anything that might turn up in their study, this principle leads to two-sided tests in most cases. The question which motivated the study should be paramount. A one-sided question should translate into a one-sided hypothesis and corresponding test.

Comparing means

Professor Hand notes that one treatment giving larger values than another does not necessarily translate into a comparison of group means. Another reason why it might not is that the effect of therapy may vary among patients. Under these circumstances, the usual two-sample tests may be very insensitive. Generalizations which deal with these problems are readily available (O'Brien, 1988).

Multiple-comparison procedures

Another example where multiple univariate analyses are appropriate is when a study is conducted to test separate hypotheses regarding the pairwise comparison of multiple therapies. Physicians are understandably bewildered when told that they are not permitted to ask whether treatments A and B differ unless a global analysis-of-variance p-value is less than 0.05. Similarly, the statement that the comparison between A and B would have been statistically significant if only the investigator had not collected data on any additional treatments violates common sense.

Definition of end points in clinical trials

It is generally acknowledged that end points in a well-designed clinical trial should be accurate, reproducible, objective and quantitative. However, the overriding concern should be clinical relevance.

Multiple end points

How should the data be analysed in a clinical trial when efficacy is measured by multiple patient characteristics? As alluded to by Professor Hand, some overall measure is needed. Various methods have been proposed (O'Brien, 1984). Comparisons among procedures often focus on power. However, in transforming a multivariate observation to a univariate test statistic, the issue of clinical relevance is paramount. Should the end points be weighted equally (rank sum approach), according to the precision with which they are measured (generalized least squares) or according to which one showed the greatest difference between treatment arms (Bonferroni correction)? The answer to these questions will depend on the nature of the question which motivated the study.

Henry Rouanet (Université René Descartes, Paris): I am very pleased to applaud Professor Hand's brilliant demolition, since for many years I have been engaged in a similar undertaking. His achievement encourages me to throw my own ninth stone after the eight that he has already thrown for us.

Example 9: negligibility paradox

Let us return to comparing two treatments, this time assuming an interval scale, and consider the research hypothesis of a *negligible effect*, stating that the difference δ of population means, even though not exactly null, is sufficiently small to be ignored (in the jargon of pharmacologists, the two drugs are 'bioequivalent'). Now suppose that two experiments I and II have been performed, both leading to the same observed effect d of very small magnitude, in that not only d but also $2d$ are deemed to be small; suppose further that in experiment I the t-test comes out to be non-significant, with $p=0.50$ (two sided), whereas in experiment II it is just significant at $p=0.05$ (two sided). Which of the two experiments is more in favour of a negligible effect δ? This situation is paradoxical, because, since the null hypothesis $\delta=0$ is consistent with I and not with II, we are tempted to say that experiment I is more in favour of a negligible effect.

Let us now recast the research question in Bayesian terms. If ϵ denotes some value deemed to be small (such that $2|d| \leqslant \epsilon$), for which experiment is the probability $P(|\delta| < \epsilon)$ greater? A Bayesian analysis, assuming a non-informative prior, shows that $P(|\delta| < \epsilon)$ is greater for experiment II; for instance, for $\epsilon = 2|d|$, we find that $P(|\delta| < \epsilon)$ is about 0.73 for experiment I, and 0.975 for experiment II.

My point with this negligibility paradox is that, to improve communication between statisticians and researchers, 'increasing the precision' of questions may not suffice, and that *changing the statistical framework*, namely, here, shifting from hypothesis testing to Bayesian inference, may constitute a big step forwards.

If researchers find it so difficult to formulate their objectives better, it is also perhaps because they are hindered by the 'basic' statistical techniques they have been taught. Specifying a significance level in advance, choosing a one-tailed *versus* a two-tailed test, etc., all such options, far from providing the first steps of a genuine statistical strategy, too often behave as traps or dead ends. I look forward to learning soon that the deconstruction enterprise is being extended to the teaching of statistics.

T. M. F. Smith (University of Southampton): I have much sympathy with the position taken by Professor Hand. In particular I agree with him that the questions addressed by the formal theory of statistics as currently taught are rarely those of direct interest to applied scientists. The p-values that we add to data analyses provide only pseudo-respectability; they rarely contribute to scientific discovery. However, his suggestion that to address the right questions we should teach within a substantive context has only limited appeal. In the past this approach has led to the fragmentation of statistics into econometrics, psychometrics, chemometrics, etc. with few obvious benefits. Also we must recognize that the main source of trained statisticians has been, and is likely to remain, the mathematical sciences. Which substantive context should we adopt when teaching mathematicians? The reality is that most mathematics students have only a limited background in any form of science and they lack the motivation and time needed to study a different discipline in sufficient depth to be able to ask relevant questions. It does take a long time to become a useful statistician!

What should we teach to mathematical scientists so that they can make a distinctive contribution to applied science? The basic concepts are probability, variability, populations and samples. The theoretical framework is that of probability and this must be taught thoroughly. Probabilistic thinking is part of our scientific heritage and together with applied probability modelling gives a key role for the statistician. Variation and sampling follow naturally and lead to issues of design. Professor Hand rightly accents the impact of the computer on the teaching of inference. We need no longer be constrained by analytical convenience. Sampling distributions can be simulated from complex and realistic processes, from sequences rather than cross-sections. We should follow Deming and adopt a systems approach.

My final comment is about errors of the third kind. In the life and social sciences most theories are descriptive. Prediction limits based on empirical models have limited value, and the disaggregation effects highlighted by the paradoxes of Simpson and Lord make interpretation hazardous. Perhaps we should follow the engineers and multiply our model-based results by misspecification factors so that our statistical edifices remain in place for a longer time. Studies to evaluate these factors would be a useful contribution of statistics to applied science.

John Tukey (Princeton University): My reaction to Professor Hand's very interesting paper is very strong, yet to some it will be paradoxical. I agree most heartily with the direction and emphasis of Hand's main thrust—we *do* need to go back into the client's motivation and into the formulation of his questions—yet I disagree, almost equally heartily, with many of the ways in which he proposes to supplement this important process.

I would not, for example, try to dragoon a climatologist into giving up a comparison of the mean temperatures of London and Paris from 1500 to 1550 because the temperatures concerned were not measured on precisely an interval scale (as they could not be at those dates). The scale involved was sufficiently well determined to make the use of the arithmetic mean quite palatable, with deviations from one kind of mercury-in-glass thermometer to another almost certainly smaller in magnitude than the other ills that plague long-term weather records. The data analyst's role should be to help the client, not to coerce either client or data unnecessarily.

In the last paragraph of the third page, Hand says

> 'Model fitting involves optimizing some criterion. Some criteria have attractive theoretical properties, but all too often the criterion is adopted by default with no thought being given to its suitability to the problem at hand. Modern computer power, however, has opened up the feasibility of using any of a vast range of criteria. Different criteria have different properties and it is necessary to consider which one best matches the aims of the study.'

Our profession cannot, in the long run, afford the hubris of asserting either

(a) that there is a single answer or
(b) that we are prepared to find it.

The proper conclusion from the first half of the quotation is therefore we should expect

(a) to make more than one fit, perhaps several, and
(b) that we should then try to understand the causes of any meaningful differences in the results.

Section 3.2 begins by assuming that the scale on which z_i is expressed is not numerically useful. The last few sentences of the section indicate the opposite, something which seems to be much more frequent.

Near Table 5, it is stated that both analyses will not normally be conducted. This should be a cause for deep regret, not something to be accepted unconcernedly.

I doubt the statement just before the conclusion that 'a pragmatic study is more likely to be intrinsically multivariate'. With the usual size of sample, a univariate analysis of a composite is likely to be more precise and more understandable.

Mervyn Wise (Leiden University): An example like example 4 was an analysis by Wishart (1939) on 20 pigs fed with (extra?) protein: did they put on weight? Unfortunately I do not remember any details, such as whether there were controls.

There are many similar problems. Two chapters in a recent book are particularly interesting (so are many others in the book, for Professor Hand's whole stimulating approach), namely 'Regression towards the mean' and 'Before–after comparisons'. (Andersen, 1990).

We can start by analysing changes in weight or other variable as a function of its *mean* (Oldham, 1962, 1968). Many questions are still unanswered. A puzzling one arises because the slope (rate of change) is often assumed to be constant and is then regressed linearly on this mean value, or on the initial value. Yet nobody seems to have considered the fact that the resulting lines in the set are concurrent!

In example 3 the numbers chosen in Table 2 'to keep the example straightforward' seem impossibly unreal. For any one journey by car made on different occasions along the same route in the same car, the instant rate of fuel consumption, whether expressed per unit time or distance, must vary considerably within and between journeys. The distribution and nature of the variations would have to be explored, perhaps with the help of inverse Gaussian distributions (Folks and Chhikara, 1978) involving passage times of particles undergoing random walks with drift and reaching a threshold.

There are again many analogous problems. In medicine 'fuel' could be replaced by a therapeutic dose, or by something harmful. The dose needed to attain a threshold could be regarded as the independent variable or, conversely, this could be the response. In any case the dose rate as a function of time can take many different forms. The response, especially if continuous (not all or nothing) also needs to be defined.

I am not happy with '*the* scientific question' or with '*the* statistical hypothesis' (my italics) in such situations when there are so many possibilities.

One general comment: 'explanatory *versus* pragmatic studies' receive apt attention, but how about exploratory studies (Chatfield, 1985), especially in relation to a model as 'a *family* of mathematical descriptions', with which I heartily agree, and looking for patterns in the parameters (—or in part contrast with—Ehrenberg and Bound (1993))?

Jacques Zighera (Université Paris X, Nanterre): Professor Hand raises many interesting issues. I fully understand that his examples are wholly illustrative, but I would like to build more extensively on example 5; briefly, if age is relevant, finding a different result separately for under and over 65 years and for the whole population does not prove anything: we must 'deconstruct' further. What follows may also be read in the light of Section 5.1.

In example 5, the underlying hypotheses are that sex and age are both relevant variables for psychiatric patients; if they were not, the population would be homogeneous and we do not need to know more; if only sex was relevant, it would be enough to know that the proportion of males is slowly decreasing.

But the author reminds us that 'model fitting involves optimizing some criterion' and it should be emphasized that this applies to supposedly simple operations such as creating groups of age inside which numbers are summed or averaged.

If age is relevant, then the real question is 'if we control for age, is it significantly true that for people of a given age (or maybe better for a given cohort of people born in the same year) the proportion of males is increasing?'. This result may hold for any age and the limit of 65 years is not important, or it may hold over a certain age and it is improbable that this age will be 65 years. By choosing, independently of any optimization procedure, the 65-years limit as relevant, we are arbitrarily—meaning without optimization criterion—projecting a 100- (possible years of age) dimensional space into a two- (over and under 65 years) dimensional space. Further, for the same reasons presented in example 5, it is possible that the proportion of males may decrease for each year of age (as for the population as a whole), whereas at the same time it increases for the two age groups.

In the same vein, I have always been worried by the discussion on whether there should be systematic mammography for women over 40 or over 50 years of age. It cannot possibly be as simple as that, and with sufficiently large samples we could build an efficiency curve of mammography according to years of age, and the age where to start (probably not 40 or 50 years) should be the question asked of the statistician (and the economist).

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

I would like to express my appreciation to everyone who contributed to the discussion.

Professor Nelder distinguishes between analysis and predictive phases. Presumably, in example 6, the analysis phase is the estimation of the probabilities of depression in each cell of Table 5. This may be done by an additive model or a multiplicative model, but Nelder's argument that the former is inappropriate is compelling. And then the predictive phase is using these estimated probabilities to answer the question of whether the relationship between the two estimated probabilities in the top row of Table 5 is the same as the relationship in the bottom row. How one does this depends on what we mean by 'relationship'. We could look at differences between the estimated probabilities, as do Brown and Harris, or we could look at the ratio of the probabilities, as do Tennant and Bebbington, or we could look at odds ratios, as does Dr Jones. The three predictive approaches answer different questions—none is more 'right' or 'wrong' than any other.

Dr Greenfield and Professor Evans describe some approaches which can assist in teaching statisticians how to deconstruct research questions. I endorse these suggestions but have reservations about the time they will require. In this vein, Dr Chatfield's proposal to set incomplete or even partially wrong questions to students sounds interesting. I hope that he can be persuaded to tell us how effective it was.

I am grateful to Professor O'Brien, Professor Wise and Professor Lenz for drawing my attention to further examples supporting my thesis. In Hand (1992) I pointed out some of the dangers of aggregation, but Professor Lenz goes further and identifies some interesting economic examples. Particularly important in this regard is his notion of a 'perfect' aggregation. His conclusions are for linear relationships, but it would be interesting to see more general results. At a fundamental level, Lenz's examples illustrate some of the points about economic statistics made by Professor Fessey. Also on the subject of aggregation, Professor Zighera rightly points out that the interaction lying at the heart of example 5 arises from data which have already been aggregated, by the collapsing of the continuum of age into just two categories. The researchers would need a good reason for having done so. (I had assumed that this would be based on pre- and post-retirement age groups, so that it might be of economic relevance for planning.)

I accept Professor Finney's criticism that the paper placed too much emphasis on tests of significance.

The work described by Professor Mackay and Professor Oldford, presenting statistics as empirical problem solving, sounds exciting—and just the right thing to do. And I think that they put their joint finger on a key issue when they point out that a detailed description of context is required in all examples. Without that, statistics is reduced to an arid discipline of mere calculation—a point argued in the introduction to Hand *et al.* (1994).

Professor Smith identifies probability as providing the key role for the statistician. Some would argue that statistics is more than this—so that methods which have no probabilistic basis are also the proper domain of statistics. Smith's final point—that we should include misspecification factors in our models—is one which I entirely endorse.

I would also have permitted Professor Tukey's climatologist to use the mean of his temperatures—provided that the different measuring instruments were calibrated to the same scales, up to linear transformations and allowing for some error. Had different, non-linearly related, scales been in use then I would not have accepted this. As to which model fitting criterion should be adopted, exploring the differences arising from different criteria is surely one way to help to deconstruct the question.

I intended my point that 'both analyses will not normally be conducted' to include the implication that this was so because there were, in principle, an infinite number of analyses which could be undertaken (Chris Jones provides a third). Hence my argument is that we need to decide which of the possible analyses is the appropriate one.

If I thought that my paper was about a metastatistical level, Professor Preece has moved up to a meta-meta-level with his three questions! To discuss something we have to distinguish it from other objects in the universe. That was really my aim in distinguishing between 'researchers and statisticians' and between 'scientific and statistical questions'. At a higher level, of course statisticians are researchers, and of course statistical questions are scientific questions.

I agree with him that it is a very rare client who has but a single research question. In any case, as we are so often reminded nowadays, there are almost always earlier studies on similar issues (a point also made by Professor Ehrenberg). And, like Preece, I would be very interested in a study of what statistical questions are, and of the kind of questions that can arise. Indeed, one of the motivations for my work on *statistical knowledge enhancement systems* (Hand, 1987, 1990) was a concern that the conventional expert systems architecture was inadequate for handling the range of question types that statistical consultants may have to answer (see Hand (1989) for a description of some such questions).

If, in Professor Lunneborg's illustration, the researcher's strategy was to choose the extent of trimming to remove the extreme point, then varying the amount of trimming to ensure this for each bootstrap sample is presumably the right thing to do. But this begs the question of whether that is an appropriate strategy.

I take issue with Chris Jones's statement that 'what is needed' is an average that commutes with the reciprocal transform. Such an average will certainly mean that the two researchers will draw the same conclusions. But this average (the geometric mean in this example) merely corresponds to yet another question that the researchers might really want to answer. It is not clear to me that this particular question is the one that 'is needed'.

On his second point, perhaps I could have phrased things differently. My point was that, if the researchers knew that they wanted to compare *means* (I am taking this as given) then switching to a comparison of *medians* because one was uneasy about the distributional assumptions is invalid. Of course, if they merely wanted to study 'location' or 'typicality' in some general sense, then studying medians may well be fine.

Professor Gower makes the point that effective statistical work requires that the statistician knows something about the substantive domain in question—and that the most effective work is the result of a continuing collaboration. I agree with him and also with his implicit concern that this can result in a division of statistics according to application area, a concern also expressed by Professor Smith. Perhaps this is one reason that statistics does not have the recognition and perceived importance that it should.

Dr Stone has hit the nail on the head with his suggestion that it is important that managers should be statistically literate. Indeed, I find myself slightly puzzled by the poor social reputation of statistics, given the need for managers to have at least a rudimentary understanding of basic ideas.

Toby Lewis identifies the parallel between my discussion of matching research questions with statistical questions and the problem of performance evaluation. The questions that he raises are particularly apposite at the present time, when universities seem beset by assessments, audits and evaluations. I have, within my own university, attempted to stimulate discussion of the costs of these exercises, relative to the potential return that they might bring and to the loss that undertaking them has meant in terms of the activities that they have displaced.

In reply to Professor Ehrenberg, I was not suggesting that investigators should ignore any potentially relevant factors, but simply that, if the researcher wanted to answer a particular question, then the statistical analysis should lead to that question being answered. As to his suggestion of a three-group trial in example 1, this would certainly permit both pragmatic and explanatory questions to be answered. But my point was that, if the researcher was interested in only one of these questions, then it would be wrong to adopt a design which did not permit that question to be answered. Professor Molenaar has also pointed out that there are three potential treatments here. But design D2 still seems to me to be the only way of ensuring that the sensitizing effect of the drug is not confounded with a delay in receiving the radiotherapy.

I agree with Professor Molenaar about the strong link between example 2 and the important work by Holland and Rubin. I consider the latter to be a prime example of careful deconstruction of statistical questions. As to Molenaar's last point (and also Professor Tukey's last point), to me multivariate analysis of variance involves finding a 'combination of all indicators into a univariate scale'—that which best separates the groups in some sense.

Professor Rouanet draws attention to one of the fundamental problems of a particular school of statistical inference. No such school is without its critics, and it may be that different kinds of research questions are answered best in the frameworks of different schools.

I do not see the fact that neither client nor statistician comes unencumbered into the negotiations as a stumbling block in the path of attempting to formulate unambiguous research questions, but I agree with the Drs Lovie that preconceptions need to be examined. I like their suggestion that explicit justification for the variables and so on is required from the client. Equally, of course, the client should require that the statisticians justify their choice of methods. Professor Herzberg's quoted summary of the paper by Bode *et al.* (1949) seems to me to be a beautiful encapsulation of what every consultant statistician should strive to be.

Several of the contributors disliked my term 'deconstruction', principally because of its negative connotations and its association with a particular school of literary criticism. I take their points but suggest that having a clear name by which we can refer to the exercise will help us to think about it and to discuss it. Since most of the discussants agreed with me about the importance of the exercise, this is surely to be desired.

REFERENCES IN THE DISCUSSION

Aitchinson, J. and Brown, J. A. C. (1957) *The Lognormal Distribution*. Cambridge: Cambridge University Press.

Andersen, B. (1990) *Methodological Errors in Medical Research*. Oxford: Blackwell.

Bode, H., Mosteller, F., Tukey, J. I. and Winsor, C. (1949) The education of a scientific generalist. *Science*, **109**, 553–558.

Box, G. E. P. (1993) Scientific statistics. *RSS News*, **21**, no. 4, 1–2.

Chatfield, C. C. (1985) The initial examination of data (with discussion). *J. R. Statist. Soc.* A, **148**, 214–253.

———(1988) *Problem Solving: a Statistician's Guide*. London: Chapman and Hall.

———(1991) Avoiding statistical pitfalls. *Statist. Sci.*, **6**, 240–268.

Christiano, L. J. and Eichenbaum, M. (1987) Temporal aggregation and structural inference in macroeconomics. *Carnegie–Rochester Conf. Ser. Publ. Poly*, **26**, 63–130.

Ehrenberg, A. S. C. and Bound, J. A. (1993) Predictability and prediction (with discussion). *J. R. Statist. Soc.* A, **156**, 167–206.

Evans, S. J. W. (1991) Discussion on Improving doctors' understanding of statistics (by D. G. Altman and J. M. Bland). *J. R. Statist. Soc.* A, **154**, 255.

Fisher, R. A. (1926) The arrangement of field experiments. *J. Mnstry Agric.*, **33**, 503–513.

Fisher, W. D. (1962) Optimal aggregation in multi-equation prediction models. *Econometrica*, **30**, 744–769.

Folks, J. L. and Chhikara, R. S. (1978) The inverse Gaussian distribution and its statistical application—a review (with discussion). *J. R. Statist. Soc.* B, **40**, 263–289.

Garfinkel, H. (1967) *Studies in Ethnomethodology*. Englewood Cliffs: Prentice Hall.

Gephart, R. P. (1988) *Ethnostatistics: Qualitative Foundations for Quantitative Research*. Newbury Park: Sage.

Gower, J. C. and Payne, R. W. (1987) On identifying yeasts and related problems. In *The Statistical Consultant in Action* (eds D. J. Hand and B. S. Everitt). Cambridge: Cambridge University Press.

Hand, D. J. (1987) A statistical knowledge enhancement system. *J. R. Statist. Soc.* A, **150**, 334–345.

———(1989) Emergent themes in statistical expert systems. In *Knowledge, Data and Computer-assisted Decisions* (eds M. Schader and W. Gaul), pp. 279–288. Berlin: Springer.

———(1990) Practical experience in developing statistical knowledge enhancement systems. *Ann. Math. Artif. Intell.*, **2**, 197–208.

———(1992) Microdata, macrodata, and metadata. In *Computational Statistics* (eds Y. Dodge and J. Whittaker), vol. 2, pp. 325–340. Heidelberg: Physica.

Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J. and Ostrowski, E. (1994) *A Handbook of Small Data Sets*. London: Chapman and Hall.

Hemelrijk, J. (1958) Statistical designs: proof and detection (in Dutch). *Statist. Neerland.*, **12**, 111–118.

Holland, P. W. (1986) Statistics and causal inference (with discussion). *J. Am. Statist. Ass.*, **81**, 945–970.

Hoy, D. (1985) Jacques Derrida. In *The Return of Grand Theory in the Human Sciences* (ed. Q. Skinner). Cambridge: Cambridge University Press.

Kirchgässner, G. and Wolters, J. (1992) Implications of temporal aggregation on the relation between two time series. *Statist. Pap.*, **33**, 1–19.

Lane, P. W. and Nelder, J. A. (1982) Analysis of covariance and standardization as instances of prediction. *Biometrics*, **38**, 613–621.

Lütkepohl, H. (1987) *Forecasting Aggregated Vector ARMA Processes*. Berlin: Springer.

MacLay, A. L. (1991) *A Dictionary of Scientific Quotations*, p. 34. Bristol: Hilger.

Mainland, D. (1964) *Elementary Medical Statistics*, 2nd edn. Philadelphia: Saunders.

Molenaar, I. W. (1988) Formal statistics and informal data analysis, or why laziness should be discouraged. *Statist. Neerland.*, **42**, 83–90.

Morgenstern, O. (1963) *On the Accuracy of Economic Observations*. Princeton: Princeton University Press.

Moser, C. (1980) Statistics and public policy. *J. R. Statist. Soc.* A, **143**, 1–31.

Nair, V. N. (ed.) (1992) Taguchi's parameter design: a panel discussion. *Technometrics*, **34**, 128–159.

Nelder, J. A. (1977) A reformulation of linear models (with discussion). *J. R. Statist. Soc.* A, **140**, 48–76.

———(1982) Linear models and non-orthogonal data. *Util. Math.* B, **21**, 141–151.

———(1993) Statistical packages and unbalanced data. *Comput. Statist. Data Anal.*, **16**, 403–406.

O'Brien, P. C. (1988) Comparing two samples: extensions of the *t*, rank-sum, and log-rank tests. *J. Am. Statist. Ass.*, **83**, 52–61.

———(1984) Procedures for comparing samples with multiple endpoints. *Biometrics*, **40**, 1079–1087.

Oldham, P. D. (1962) A note on the analysis of repeated measurements on the same subject. *J. Chron. Dis.*, **15**, 969–977.

———(1968) *Measurement in Medicine*. London: English Universities Press.

Prais, S. J. and Houthakker, H. S. (1955) *The Analysis of Family Budgets*. Cambridge: Cambridge University Press.

Preece, D. A. (1990) R. A. Fisher and experimental design: a review. *Biometrics*, **46**, 925–935.

Rubin, D. B. (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, **47**, 1213–1234.

Schneeweiss, H. (1965) Das Aggregationsproblem. *Statist. Hefte*, **6**, 1–26.

Schönfeld, P. (1979) Fehlspezifikation dynamischer Modelle durch temporale Aggregation. In *Empirische Wirtschaftsforschung, Festschrift für R. Krengel* (eds J. Frohn and R. Stäglin), pp. 253–266. Berlin: Duncker and Humblot.

Sondermann, D. (1973) Optimale Aggregation von grossen Gleichungssystemen. *Z. Natnolökon.*, **33**, 235–250.

Sprent, P. (1993) *Applied Nonparametric Statistical Methods*, 2nd edn. London: Chapman and Hall.

Stone, R. (1993) The assumptions on which causal inferences rest. *J. R. Statist. Soc.* B, **55**, 455–466.

Tiao, G. C. (1972) Asymptotic behaviour of temporal aggregates of time series. *Biometrika*, **59**, 525–531.

Werner, H. J. (1982) On the temporal aggregation in discrete dynamical systems. *Lect. Notes Control Inform. Sci.*, **38**, 819–825.

Wishart, J. (1939) Statistical treatment of animal experiments (with discussion). *J. R. Statist. Soc.*, suppl., **6**, 1–22.