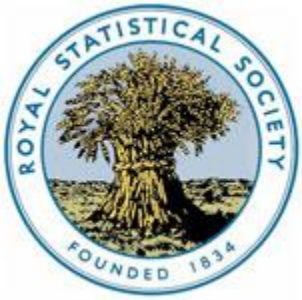

WILEY

Statistics and the Theory of Measurement
Author(s): D. J. Hand
Source: *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, Vol. 159, No. 3 (1996), pp. 445-492
Published by: Wiley for the Royal Statistical Society
Stable URL: http://www.jstor.org/stable/2983326
Accessed: 23-01-2018 20:55 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://about.jstor.org/terms

Royal Statistical Society, *Wiley* are collaborating with JSTOR to digitize, preserve and extend access to *Journal of the Royal Statistical Society. Series A (Statistics in Society)*

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

J. R. Statist. Soc. A (1996)
159, *Part* 3, *pp.* 445–492

Statistics and the Theory of Measurement

By D. J. HAND†

The Open University, Milton Keynes, UK

[*Read before* The Royal Statistical Society *on Wednesday, March 20th, 1996, the President,* Professor A. F. M. Smith, *in the Chair*]

SUMMARY

Just as there are different interpretations of probability, leading to different kinds of inferential statements and different conclusions about statistical models and questions, so there are different theories of measurement, which in turn may lead to different kinds of statistical model and possibly different conclusions. This has led to much confusion and a long running debate about when different classes of statistical methods may legitimately be applied. This paper outlines the major theories of measurement and their relationships and describes the different kinds of models and hypotheses which may be formulated within each theory. One general conclusion is that the domains of applicability of the two major theories are typically different, and it is this which helps apparent contradictions to be avoided in most practical applications.

Keywords: CLASSICAL MEASUREMENT; MEASUREMENT THEORY; OPERATIONAL MEASUREMENT; REPRESENTATIONAL MEASUREMENT; STATISTICAL MODELS; TRANSFORMATIONS

1. INTRODUCTION

No modern statistician can be unfamiliar with the fact that there are different interpretations of probability, that these lead to different schools of inference and that the conclusions drawn by these schools can differ. The dialogue about which is the 'correct' interpretation of probability has been long and often bitter, and is well documented in the statistical literature. Less well appreciated within the statistical community, however, is that there has also been a long and often equally acrimonious dialogue about the interpretation of measurement. It is curious that this debate has barely figured in the statistical literature since it is just as central to statistical work as the interpretation of probability. Instead it has occurred mostly in the social and behavioural science literature (for example, Stevens (1946, 1951), Lord (1953), Adams *et al.* (1965), Gaito (1980), Townsend and Ashby (1984), Michell (1986) and Stine (1989)). An exception to this concerns the measurement of probability itself, especially the measurement of subjective probability, which statisticians have explored in detail. However, since the aim of this paper is to draw the attention of statisticians to issues arising from measurement in general, the particular problems of measuring probability are not discussed. (A similar comment applies to utility.) Bernardo and Smith (1994) gave a comprehensive overview of the issues associated with measuring probability.

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

© 1996 Royal Statistical Society 0035–9238/96/159445

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

As with probability, different interpretations of measurement can lead to different consequences for inference — and hence to different conclusions by virtue of the fact that different kinds of hypotheses must be formulated and models built. Confusion between these different kinds of hypotheses and models has led to confusion and considerable debate over the validity of the conclusions drawn. Much of this debate has been stimulated by controversy over the legitimacy of applying different classes of statistical methods to data arising from different kinds of measurement activity. However, there is a complex interplay between transformations of variables, measurement theory, the precise nature of the research question being investigated and the meaningfulness of the results. Given the ubiquity of transformations in modern statistical analysis and the subtlety of formulating precise research questions, it is even more surprising that the statistical literature does not contain an extensive discussion of the meaning of measurement. This paper seeks to make a small step towards filling that gap.

In a study of statistical consultancy, van den Berg (1991) found that measurement level was the aspect on which there was least agreement. She reported many different classifications of measurements: counts *versus* measurements; nominal, ordinal and numerical; dichotomous as a separate category; qualitative *versus* quantitative; qualitative levels sometimes being called categorical or non-numerical; quantitative levels sometimes being called metric, numerical or simply measurement. Nelder (1990) described various 'modes' of data, distinguishing continuous counts, continuous ratios, count ratios and categorical, with the last being divided into three subtypes (nominal, ordered on the basis of an underlying scale and ordered without an underlying scale). Bartholomew (1987) distinguished metrical from categorical. Mosteller and Tukey (1977) identified grades, ranks, counted fractions, counts, amounts and balances. And other classifications have also been suggested.

Many, if not most, of these classifications are based on pragmatic data analytic grounds: the statistical techniques used to analyse a mere classification are different from those used to analyse a 'continuous numeric' variable such as length. However, as we shall see, a deeper and more fundamental classification also exists.

One of the earliest explicit formulations of measurement was due to the physicist Campbell (1920), who described ('fundamental') measurement as the assignment of numerals to represent the properties of objects, where the objects satisfied

(a) an order relationship and
(b) a physical process of 'addition' (nowadays called *concatenation*, such as placing rods end to end in a straight line).

But not all measurements in physics can be so described (e.g. density) so the basic notion had to be extended to include 'derived' measurement: those properties which are defined in terms of others. However, even this extension fails for disciplines such as psychology, where often not only can no concatenation operation be defined but also it may not be obvious precisely what empirical relationship is being represented.

In an effort to resolve this problem, the psychophysicist Stevens (1946, 1951) made two advances. Firstly, he generalized Campbell's concatenation structures to other empirical systems (so, for example, systems which satisfied ordinality but not concatenation also constituted a kind of measurement). And, secondly, he noted that the mapping from the empirical system to the numerical system did not uniquely

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

characterize the numbers to be assigned. We can, for example, measure length in inches or centimetres—both representations are equally legitimate. Given this, Stevens argued, we should use only those statistics which are invariant to changes between legitimate representations. Adopting this principle, he defined the now famous *nominal*, *ordinal*, *interval* and *ratio scales*, characterized by the type of transformations which mapped from one legitimate representation to another. Such transformations are nowadays called *admissible* or *permissible* transformations. Thus, for nominal scales one-to-one transformations are permissible. For ordinal scales monotonic increasing transformations are permissible. For interval scales linear transformations are permissible. And for ratio scales only similarity transformations are permissible.

Stevens's arguments seem sound: if statistical methods yield conclusions that vary according to which of the equally legitimate numerical representations is adopted, then surely something must be amiss. However, not everyone agreed: as some pointed out, statistical procedures make merely distributional assumptions, not assumptions about the type of scale. Statistical operations can be carried out on numbers no matter what the origin of those numbers is. Somehow the issue is deeper, involving notions of *interpretability* of data and statistical conclusions. The debate about the relationship between statistical techniques and measurement scales has continued, right from the time of Campbell and Stevens to the present. Recent contributions include Velleman and Wilkinson (1993a, b), Hand (1993a) and Niederée (1994).

Although most of the debate has taken the form of polemics favouring one or other side of the debate, a few authors have attempted to resolve things by considering higher level issues. They have suggested that perhaps the existence of more than one theory of measurement lies at the root of the controversy. Dawes and Smith (1985), for example, contrasted *representational* and *non-representational* measurement, and Michell (1986) (developed further in Michell (1990)) contrasted *representational* theory, *operational* theory and *classical* theory. Michell, in fact, described the controversy over scales of measurement as a 'clash of paradigms'—making the parallel between probability and measurement again striking.

In this paper, to provide the necessary background, the different theories are outlined in Section 2. The representational paradigm, described in Section 2.1, is by far the best developed theoretically and might perhaps be regarded as representing the current dominant paradigm (occupying the role that the frequentist interpretation of probability did a couple of decades ago?). Certainly, it seems to be regarded as having the soundest conceptual basis: when the validity of a statistical analysis and conclusion is criticized on measurement theoretic grounds, the representational theory is typically the theory referred to.

The non-representational theory described by Dawes and Smith (1985) seems to have very similar content to the operational theory described by Michell (1986) and also to the *pseudopointer* measurement of Suppes and Zinnes (1963). The ideas underlying these theories are outlined in Section 2.2, where we have adopted the term operational, partly to avoid the awkwardness of non-representational and partly so that we can outline Michell's third theory. The importance of operational measurement hinges on its place in justifying analytic practices that might be regarded as dubious under the representational theory. Operational theory had its genesis in physics, where there was an uneasiness about the reality of the theories and

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

concepts introduced in the first decades of this century, but it seems to have had most practical effect in disciplines such as psychology. In particular, it has achieved a high level of sophistication through statistical models such as latent variable models and linear structural relational models. Finally, in addition to the representational and operational schools, Michell describes a *classical* school; this is outlined in Section 2.3.

One way of thinking about the difference between representational and operational measurement is that the former seeks to *represent* or *model* empirical relationships — and so is about *understanding* the substantive domain of investigation — whereas the latter seeks to *predict*. Accurate prediction can be achieved without any understanding of the underlying mechanism (witness someone who can drive a car well without any understanding of how it works). Confusion between these two aims is widespread and is probably promoted by the unfortunate adoption within statistics of the term 'model' to denote a mere description. Things would be clearer if model were reserved for a system which reflected a theory or hypothesis about an underlying mechanism, and if some other term (such as 'description') were used for a summary of the data which, though perhaps well fitting, was no more than an empirical construct. However, the momentum of statistical usage is too great for me to attempt to change it here, so, where it is important in what follows, I shall refer to a 'mechanistic model' and a 'descriptive model' as appropriate.

In Section 3.1 I examine the formulation of statistical models (subsuming both types) and hypotheses in detail, adopting strands from both the representational and the operational schools. I examine the concept of *meaningfulness* of statistical statements, primarily from Stevens's perspective of invariance over legitimate representations but also briefly from the perspective of definability in terms of the system being measured. I argue that statistical statements about samples or groups of objects *need not* be defined *solely* in terms of the empirical relationships between objects. To a large extent, whether or not they should be so defined depends on whether one is building a mechanistic or descriptive model. I hope that this sheds some light on the controversy over what and when statistical operations are legitimate.

In Section 3.2 I look more closely at transformations and their relationships to measurement scales. It may be possible to find numerical assignments, each satisfying certain properties which remain invariant under the same classes of transformations, but which are not homomorphic to each other. Superficially this seems to suggest that we have alternative non-homomorphic representations for a given empirical system, suggesting that Stevens's strictures are too severe. Closer examination, however, shows that this is not so — the representations reflect different aspects of the empirical system.

Section 3.3 examines the relationship between model generation, model testing and the underlying measurement theory. Model generation is the more relaxed activity, in that, almost by definition, a search for unexpected pattern should not be constrained by pre-existing views on what is and is not legitimate. Model building and model testing, however, are necessarily constrained by the properties of the data with which we are dealing: one's philosophy of measurement influences the hypotheses that one can formulate and the ways in which those hypotheses may be examined. The detailed formulation of a statistical question determines what kind of invariances are required of the data for it to be meaningful, and some authors have taken this to

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

define the measurement scale of the data. In my view, however, this leads to needless complications.

There are aspects to the relationship between measurement and statistics that are beyond those covered in this paper, although inevitably intertwined with them. An important one, merely mentioned below in the context of the sort of underlying theories that should be built to model it, is accuracy of measurement. Bailar (1985) presented a broad overview of measurement accuracy, covering the notions of replication and reproducibility, and giving examples from a wide variety of areas. Closely related, but, as statisticians will be aware, not identical, is the concept of precision. Wise (1995) provided a collection of essays focusing on the historical development of the concept. A third, broader, aspect of the relationship between the two disciplines is the question of what to measure. Sometimes we shall choose to measure a proxy variable because the thing that we are really interested in is difficult or expensive to measure, or because it defies clear definition (see Section 2.2). The proxy variable may be well defined, it may be based on sound statistical principles and it may be precise, all features which make it attractive—but using it as, for example, a control mechanism may lead to distortion of the objectives. Topical examples are organizational audit mechanisms such as educational league tables and the research assessment exercise used in assessing research performance of British university departments.

2. THREE THEORIES OF MEASUREMENT

2.1. *Representational Measurement Theory*

Representational measurement theory is the dominant current measurement paradigm. In fact, the phrase 'measurement theory' is often used as a shortened term for this particular theory. It originated around the end of the 19th century and the beginning of the 20th century with the work of von Helmholtz (1887) and Hölder (1901). The *magnum opus* of representational measurement theory is the three-volume work *Foundations of Measurement* (Krantz *et al.*, 1971; Suppes *et al.*, 1989; Luce *et al.*, 1990). The approach adopted by this work, and by representational measurement theory in general, is made clear in the opening sentences of the preface:

> 'Scattered about the literature of economics, mathematics, philosophy, physics, psychology, and statistics are axiom systems and theorems that are intended to explain why some attributes of objects, substances, and events can reasonably be represented numerically. . . . Although such systems are of some mathematical interest, they warrant our attention primarily as empirical theories—as attempts to formulate properties that are observed to be true about certain qualitative attributes.'

So, representational measurement theory is about describing real empirical systems.

In representational measurement theory we begin with a set of *objects*, each of which has one or more common *attributes*, each in turn of which can be divided into mutually exclusive and exhaustive equivalence classes. To keep things simple at this stage, we restrict the discussion to a single attribute. Thus each object can be uniquely allocated to a single equivalence class according to the 'value' of its attribute. Then the objects and the relationships between them (induced by the relationships between the equivalence classes for the attribute) constitute an *empirical relational system* (ERS). In parallel with this we construct a *numerical relational*

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

system (NRS) comprising numbers (typically the real numbers, though they need not be) and the relationships between them. Then representational measurement theory is concerned with establishing a mapping from the objects, via the equivalence classes to which they belong, to the number system in such a way that the relationships between objects are matched by relationships between numbers. These numbers form the values of a *variable*. In particular, representational measurement theory presents axioms which the objects must satisfy to permit such numerical representation.

Statistical operations can then be carried out on the numbers and the aim is that conclusions reached about relationships between the numbers will reflect corresponding relationships between the objects. In particular, in statistical analysis we may be interested in making *inferential* statements about notional *classes* of objects, from which those actually studied were drawn.

A simple (and basic) example will illustrate these ideas.

Consider a set of rigid rods. These have an attribute 'length' and we can allocate rods to equivalence classes according to whether or not, when laid side by side so that the left-hand ends terminate together, the right-hand ends also terminate together. All those with both ends terminating together constitute a single class. But we can say more than this. These classes are related according to whether the rods in one class terminate to the right of the rods in the other class. When this is satisfied we say that the rods in the first class 'are not shorter than' the rods in the second.

We can now establish a mapping from the rods to the positive real numbers such that longer rods are associated with larger numbers, i.e. letting $M(x)$ be the number corresponding to rod x, we assign numbers such that $M(x) \geqslant M(y)$ if and only if $x \supseteq y$, where $\supseteq$ represents 'is not shorter than'. $M(x)$ is the value of the *variable* 'length' for the rod x.

By these means we establish an isomorphism between the equivalence classes and the positive real numbers, and a homomorphism between the rods and the positive real numbers. In mathematical terms, we establish a homomorphism from the ERS denoted by $[A, \supseteq]$, where A represents the set of rods, to the NRS denoted by $[R^+, \geqslant]$. An important feature of this procedure is that, in general, the homomorphism will not be unique — there will be more than one mapping in which the relationships between the numbers reflect the relationships between the objects. Now, given a set of numbers which have been assigned to the objects in such a way that they preserve the is not shorter than relationship, we can carry out statistical operations on those numbers, using the $\geqslant$ relationship, and any conclusions that we reach will have empirical counterparts. We could, for example, compare the medians of the lengths of two groups of rods.

As it happens, we can go further with this example. The attribute length possessed by rigid rods has other internal relationships. In particular, if we place two rods end to end in a straight line then we can (in principle, at least) find a third rod which, if placed next to this concatenation, has left- and right-hand ends aligned with those of the concatenated pair. We thus have a three-component relationship between the rods in addition to the two-component relationship above. Such three-component relationships are often written in *operation* form as $x \circ y = z$, symbolizing that $x \circ y$, the concatenation of x and y, has aligned ends with the single rod z. It turns out, not surprisingly to anyone steeped in Western culture, that we can find an NRS which, in addition to reflecting the relationship $\supseteq$, also reflects the relationship $\circ$. In particular, we can represent the relationship $\circ$ by $+$. Thus we can assign numbers to the rods to

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

represent their lengths such that $M(x \circ y) = M(x) + M(y)$. Mathematically, we can find a homomorphism M from the ERS $[A, \supseteq, \circ]$ to the NRS $[R^+, \geqslant, +]$. Now we can undertake statistical operations which include addition of the numbers—such as comparing the total lengths of two groups. Again such operations will have empirical counterparts. Note that the introduction of the extra relationship, $\circ$, reduces the set of mappings which preserve the relationships.

The rods example is fundamental: length and weight measurement were used by Campbell (1920) to illustrate what he called *fundamental* measurement. They are examples of what are called *extensive* or *additive* measurements since the concatenation operation (e.g. placing two weights together on the same pan of a weighing balance) can be directly represented by addition.

In general, for a set of objects A, if the attribute has relationships $R_1, R_2, \ldots, R_n$, then we seek to establish a homomorphism from the ERS $[A, R_1, R_2, \ldots, R_n]$ to an NRS $[R, r_1, r_2, \ldots, r_n]$, where the r_i are relationships between numbers. Different relationships R_i will be represented by different r_i. In general, we can produce axioms that the empirical system must satisfy to permit representation by a given numerical system. For example, axiom systems for extensive structures, introduced by Hölder (1901), are given by Pfanzagl (1959), Suppes (1951), Suppes and Zinnes (1963) and Narens and Luce (1986) and have been generalized in various ways (e.g. Roberts and Luce (1968) and Narens (1974)).

Now, as noted above, the homomorphisms from the given ERS to a particular NRS will not, in general, be unique. There will typically be more than one set of numbers which models the empirical relationships, so that the R_i may be accurately represented by r_i for more than one numerical assignment. For example, given an acceptable assignment of numbers to the lengths of the rods in the above example, then an arbitrary rescaling of the lengths (changing inches to centimetres, for example) will also produce an acceptable assignment: the ordering and the end-to-end concatenation operation $\circ$ will be properly represented by $\geqslant$ and $+$ respectively, in both numerical assignments. More generally, the *structure* of a model must be invariant to changes in the numerical assignment. This is what lies at the heart of dimensional analysis in physics—so that, for example, changing the units in which length is measured leads to balancing changes on both sides of a model formula. The dimensions of length must be balanced. Finney (1977) pointed out that dimensional analysis is at least as applicable to statistical models, presenting a series of examples which show how the method can be used to detect model inadequacies.

The fact that a given relationship between objects can be represented by a particular NRS in more than one way induces a taxonomy on the representations—and hence leads to the notion of *types* of scale. The set of homomorphisms leading to numerical representations of the ERS and which are related by a given type of transformation fall into one class. Those related by another type of transformation fall into another class. And so on. This is also the essence of Stevens's (1946, 1951) scale types. In fact the modern classification is produced by noting that there is a one-to-one correspondence between the set of homomorphisms of an ERS into an NRS and the group of automorphisms of the ERS and then classifying the automorphism groups. The latter is done in terms of the *degree of homogeneity* (k) and the *degree of uniqueness* (l) (Narens, 1981a; Narens and Luce, 1986) of the ERS. These tell us the size of structures which are preserved by the automorphisms. Using the pair (k, l) to classify scales, we find that ratio scales are of type (1, 1), interval scales are of type

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

(2, 2) and ordinal scales are of type (∞, ∞). Moreover, various results have also been established about the possible scale types that can arise—so helping to explain why so few scale types are used in the sciences. Details were given by Narens (1981a, b), Luce and Narens (1983, 1985) and Alper (1984, 1985, 1987).

Although concatenation operations, yielding extensive measurement, played a fundamental role in the early development of formal measurement theory, and are central to the physical sciences, they are of little use in the social and behavioural sciences where concatenation operations are typically unavailable. As mentioned above, this absence has been the source and stimulus of much of the work on measurement theory. It stimulated thought about alternative theories, as outlined in Sections 2.2 and 2.3, which was at the root of the controversy mentioned in Section 1, and led to the development of alternative axiomatic structures which have subsequently also become important. These include models for forming weighted means (e.g. of expected utility, by Von Neumann and Morgenstern (1947)) and *conjoint measurement*. The latter development dates from the 1960s and produces interval scales solely from an ordinal starting point. For example, suppose that we have three attributes R, X and Y, such that for each pair (X, Y) there is a unique corresponding value of R. Then, given certain restrictions on the empirical system (such as a condition which can be loosely interpreted in statistical terms as there being no interaction between X and Y in their effect on R), numerical assignments r, x and y can be made to R, X and Y such that $r(x, y) = x + y$. Important early references are Krantz (1964), Luce and Tukey (1964) and Holman (1971). Working independently, Rasch (see, for example, Rasch (1977)) showed that the existence of order preserving numerical representations r, x and y of R, X and Y such that $r(x, y) = x + y$ led to interval scales for comparing the Xs (and Ys).

The fact that the homomorphism from the ERS to the chosen NRS will generally not be unique should not be confused with the fact that R_i may be representable by a different r_i. For example, the addition of numerical length measures in the rods example can be replaced by multiplication of exp(those numerical length measures). Representations in which addition is the numerical operation are by far the most common, but they are not the only ones. Electrical components placed in parallel can be measured in terms of conductance, in which case addition of the values is appropriate. Or they can be measured in terms of resistance, in which case the appropriate numerical operation corresponding to 'parallel concatenation' is $\rho_1 \oplus \rho_2 = (\rho_1^{-1} + \rho_2^{-1})^{-1}$. Similarly, velocities may be mapped to the NRS $[R, \geqslant, +]$ in the usual classical way, or they can be mapped to $[(0, c), \geqslant, ®]$ where ® is relativistic combination of velocities, given by

$$u®v = \frac{u + v}{1 - uv/c^2}.$$

In each such mapping, since the same R_i is being represented by the various r_i, the numerical representations must be isomorphic. In these three examples the isomorphisms are given by the respective transformations $\exp x$, $1/x$ and $\tanh^{-1}(x/c)$.

Of course, care must be taken to be sure that the various alternative representations describe the same empirical operation: electrical components placed in series *are* additively represented in terms of resistance. An example of an unusual concatenation operation which may appeal to statisticians is the following alternative

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

concatenation for rigid rods (Ellis, 1966). Instead of placing two rods a and b end to end in a straight line, concatenate them by placing them end to end at right angles. The third rod, of length equal to the concatenation of the first two, $a \circ b$, forms the hypotenuse of the right-angled triangle. Can we find a numerical assignment which will properly represent the lengths and in which addition of numbers corresponds to this concatenation operation? If so, how is this assignment related to the conventional assignment for end-to-end concatenation? The answer to the first question is that we can, and the answer to the second is that numerical lengths arising from the right-angled concatenation are the squares of lengths arising from the end-to-end concatenation. A ruler graduated using squared numbers in place of the conventional graduations would yield a perfectly legitimate alternative description of space, but one in which summation of lengths corresponded to right-angle concatenation in place of end-to-end concatenation. At first this may seem to lead to a horribly contrived description of Euclidean space, but it is precisely what is used in statistics when variations arising from multiple sources, described in terms of variances (in squared units of measurement), are added.

2.2. *Operational Measurement Theory*

Operationalism defines scientific concepts in terms of the operations used to identify or measure them. It avoids assuming an underlying reality and so is fundamentally different from representationalism, which is based on a mapping from an assumed underlying reality. In operationalism, things start with the measurement procedure. Operationalism was developed by Bridgman (1927) and adopted by Dingle (1950), who summarized it thus:

> 'Formerly science was regarded as the study of an external world, independent of the observer whose experiments and observations were simply means of finding out how the world was constructed and by what laws its behaviour was governed. The emphasis has now shifted from the nature of the world to the operations of experiment and observations. These are no longer regarded as more or less arbitrary means of discovering the already established order of nature, but rather as affording primary data for rational study; and any world that we may contemplate is no longer an independent existence whose nature demands or determines them, but rather a logical construct, formed and shaped and modified so as to afford a true picture of the relations which the observations exhibit.'

Thus, an attribute is defined by its measuring procedure, no more and no less, and has no 'real' existence beyond that. In operationalism the attribute and the variable are one and the same. This approach thus defines 'a measurement [as] any precisely specified operation that yields a number' (Dingle (1950), p. 1).

It follows that, to be useful, the numerical assignment procedure has to be well defined. Arbitrariness in the procedure will reflect itself in ambiguity in the results. This is one reason why problems arise in the social and behavioural sciences, where, inevitably, measuring procedures are complex. A complete specification of the procedure is often difficult or impossible and different researchers may use the same name for variables that actually have subtly different definitions, leading to different conclusions. Since the definition of the concept lies in the measurement procedure it is not a cause for concern that different procedures lead to different conclusions, but rather an indication that more refined theory needs to be developed.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

This issue of slightly different definitions for a given variable name in operational theory is sometimes confused with the fact that there may be different ways of measuring an attribute in the representational theory. We can, for example, measure length by using rigid rods or by using the time for light to transit from one point to another. These can be regarded as operational definitions of different length concepts (which empirical study shows to be very highly correlated) or, by virtue of the complex physical theory which has been constructed, to be different descriptions of the same underlying attribute. In contrast, it is easier to accept that two questionnaires, seeking to tap into some attitude to a proposition but using different questions, are describing slightly different phenomena. If, in a representational model, two supposedly alternative measurement methods lead to consistently different results, then this is an indication that the ERS is not as straightforward as was thought. The distinction between mass and weight provides an illustration.

Niederée (1994), p. 568, said of the operational approach, which removes ambiguity by defining a phenomenon in terms of a specified measurement procedure:

> 'This outspoken conventionalist procedure appears suitable for bureaucrats, say, who just want to establish plausible formal decision rules, or for practical situations where "it doesn't really matter", . . . or in scientific contexts where some vaguely formulated theory is to be rendered plausible with the help of generally accepted statistical methods. . . . But in many scientific or practical contexts, this strategy usually just begs the question.'

I agree that in many applications such an approach may not be suitable. However, in others it may be. Firstly, if everyone uses the same conventions to discuss some phenomenon then useful discussions can take place. Bureaucrats are not the only people for whom this is necessary. And, secondly, operational measurements in which the measurement sits properly and effectively in a theoretical web of relationships with other variables—i.e. those which yield effective *predictions*—are useful. (Non-useful measurements presumably are not used, at least in good science.) This is made apparent by the notion of *construct validity*, which assesses the extent to which the measure conforms with the theoretical predictions of relationships with other variables. (Of course, lack of construct validity may mean that the measure is a poor measure of the theoretical concept in question, but it could also mean that the theory relating it to other variables is inadequate or that these other variables are poorly measured. But that is another issue.) If measurements in the physical sciences are viewed in operational terms then they provide examples of very high construct validity: the theoretical predictions conform very closely with measurement outcomes.

The related concept of *criterion-related validity* refers to the accuracy with which the measurement procedure predicts an external criterion (such as true ages in a study of subjective age assessments or a 'gold standard' in medicine). Criterion-related validity is probably more relevant in a representational than in an operational context where there is a clear objective criterion that we are trying to predict, namely the behaviour of the ERS.

Some people find distasteful the fact that operational theory draws conclusions only about the results of measurement procedures, and leaves researchers to make an inference to an 'underlying reality' if they so wish. But statisticians regularly use a parallel strategy in a different context—that of randomization or permutation tests.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

These make inferences conditionally on the data, and then let the researcher make non-statistical inferences to wider populations.

Techniques for constructing operational measurements fall into two classes: those that focus on single variables and those that define a variable in terms of others. An operational definition of length in terms of laying a ruler end to end in a straight line is of the former type. Campbell's derived measurements, such as density, are of the latter type (though, of course, they may also have deeper representational interpretations in terms of fundamental measurements). Examples of the former in the behavioural sciences are paired comparisons and rating scales. Examples of the latter are Guttman scaling and unfolding methods (see, for example, van der Ven (1980)). Multidimensional scaling (Cox and Cox, 1994) is also an example of the latter, though until recently the intrinsic non-linearity of many of the methods made it difficult to reify (i.e. to give a meaning to) the dimensions of the lower dimensional representation space in terms of the contributing variables. This has now been overcome (see, for example, Gower and Hand (1996)).

Optimal scaling methods such as correspondence analysis and the more general methods described by Gifi (1990) should also be mentioned in this context. These identify a numerical coding of the raw variables which optimizes some additional criterion—typically some relationships between variables. For example, we might find that particular numerical assignment for two ordinal scales which optimizes the Pearson correlation coefficient between them, subject to fixed means and variances. Or we might find that particular numerical assignment which maximizes the minimum possible correlation between the assigned numbers and all possible patterns of numbers satisfying constraints such as ordinality (as is explored by Abelson and Tukey (1959, 1963)). Such statistical techniques identify a particular mapping from the objects to numbers, i.e. they identify a unique measuring instrument. This leads us to a fundamental point: in representational theory the number assigned to an object is not unique; it could be any number from a set of numbers. However, for a particular object, the number chosen depends on the numbers chosen for the other objects—and this dependence arises via the empirical relationships between the objects. In contrast, in operational theory the number assigned to an object is unique—it emerges from the measuring instrument. Of course, for a particular statistical statement, other numerical assignments might yield the same truth values. For example, if the integers 1–10 are assigned to rocks according to their relative hardness, then the statement that $\bar{x}_1 > \bar{x}_2$ for two samples of rocks would have the same truth value if the coding 21–30 had been adopted instead. So, clearly, similarity transformations do not influence the validity of this statement. This suggests that a notion of scale type may be definable for operational measurements. We return to this in Section 3.3.

Perhaps the statistical methods that are most widely used in the behavioural sciences for constructing operational definitions of variables are latent variable models, which explain the relationships between the *observed* (or *manifest* or *indicator*) variables in terms of hypothesized unobserved (and unobservable) *latent* variables. It seems to me, however, that the so-called *latent* variables are operationally defined by their relationships to the observed variables. The term 'construct', which is also occasionally used, is a much more appropriate term. Quality-of-life scales, for example, are clearly constructs, defined in terms of their constituent components, rather than underlying variables. Price and quantity indices in

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

economics might also be regarded as constructs. One approach to defining such indices is to establish a set of axioms (to be regarded as 'self-evident') and conditions (called 'tests') which the indices must satisfy, and then to derive the forms which do satisfy them (for a review, see Balk (1995)). This leads to an axiomatic system which has parallels to the representational approach. However, whereas in representationalism the axioms describe how the empirical objects must behave to permit certain numerical representations, here the axioms describe desirable properties for the measures themselves. Thus this approach seems more naturally described as operational.

2.3. *Classical and Other Theories of Measurement*

Subjective and frequentist interpretations of probability are but two of many. Indeed, even within these labels there are different schools of thought. Similar diversity applies to theories of measurement: there are other variants in addition to the representational and operational schools. Kyburg (1984), p. 253, for example, stated:

> 'Most approaches to measurement that have been suggested in recent years have taken the process of measurement to be the assignment of *numbers* to objects and events. I have suggested that the value (or interval of values) assigned to an object or event by measurement is a magnitude (or interval of magnitudes), rather than a number.'

Thus, instead of assigning the number '2' to the length, in feet, of an object, he assigns the magnitude '2 ft' to the object. His book develops the consequences of this approach. Kyburg (1984) also drew attention to the central role of error in measurement. Measurement error represents yet another link between the two concepts of measurement and probability.

Any mismatch between theoretical predictions and observations can be explained in two ways: either the theory is inadequate or there is measurement error (or, and probably more usually, both). Implicit in representational measurement is a theory about the objects: that they are related (in terms of some kind of behaviour — the attribute) in certain ways that form the relationships of the ERS. So, for example, we might assume that the objects are ordered and satisfy some concatenation relationship, despite the fact that we can establish this only for a finite number of (sets of) objects. The question of establishing relationships for the ERS is tightly bound up with the problem of induction — the core of statistical inference itself. If the assumed relationships do not hold, or hold only approximately, then we should expect the predictions and inferences drawn from our numerical calculations not to hold or to hold only approximately. But such approximations will also appear if the measurements are not perfectly reliable — measurement error manifests itself most clearly in the fact that repeated measurements of the same attribute of the same object (using the same measuring instrument) can yield different values. It is a ubiquitous aspect of measurement in all scientific investigation. As such, one might argue, it should be integrated into the theoretical structure describing measurement. Attempts in this direction have been made by, for example, Falmagne (1979, 1980).

Michell (1986, 1990) described yet another theory of measurement — the classical theory, which he contrasted with the representational and operational theories. He called it classical because, according to him, traces may be found in the works of Aristotle and Euclid and it was

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

'developed during the Middle Ages and the Scientific Revolution and sustained the practice of measurement until at least the beginning of this century'.

According to this theory, measurement addresses the question of 'how much' of a particular attribute an object has and thus only refers to attributes which are 'quantitative'. It is this term quantitative on which this third, fundamentally realist, theory hangs.

A quantitative attribute is an attribute whose values satisfy ordinal and additive relationships — Michell distinguished this approach from the representational theory by stressing that it is the *attribute* which has these properties and not the objects. The behaviour of a set of *objects* may or may not reflect the quantitative nature of the attribute in question: the behaviour of objects is a function of their other properties as well as the attribute in question. A physical concatenation operation between objects certainly provides evidence for the quantitative nature of an attribute, but the lack of such an operation does not mean that the attribute is not quantitative. Evidence for the assertion that an attribute is quantitative may be found in other ways. Michell cited the example of temperature: objects that have temperature do not satisfy a concatenation relationship and yet this attribute is generally regarded as quantitative.

According to this theory the hypothesis that an attribute is quantitative is a scientific hypothesis just like any other. Measurement then involves the discovery of the relationship between different quantities of the given attribute. The key word here is 'discovery'. Whereas the representational theory *assigns* numbers to objects to model their relationships, and the operational theory assigns numbers according to some consistent measurement procedure, the classical theory *discovers* pre-existing relationships. By definition, any quantitative attribute has an associated variable.

Developing a measurement procedure according to the classical theory requires relating the hypothesized quantitative attributes to observable quantities within some theoretical framework. The hypothesized quantitative attributes can then be measured by virtue of their relationships. Here the hypothesized attributes, as well as their quantitative nature, are all a part of the theory being studied. Rasch's (1977) notion of *specific objectivity* might be regarded as fitting naturally into this framework. Rasch gave an example in which observed scores on a test are described by a Poisson model. The parameter of the Poisson model can be viewed as an underlying measure of ability for a given test. Rasch then showed that this model permits parameters to be separated into a set describing comparisons between the abilities of individuals (independent of which test is used) and a set describing comparisons between test difficulties (independent of which subject is assessed). (It is this separation which leads one to believe that, for example, ability is an intrinsic property of the individual.) The link between the observed scores and the parameters is

(a) stochastic and
(b) non-linear

— quite complicated, as Michell suggested it might sometimes be. (Of course, we could also regard the parameters as descriptive constructs — as in operational theory — rather than as underlying real quantitative attributes.)

As noted in the preceding section, many researchers view statistical tools such as

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

factor analysis as being ways of measuring an underlying latent variable via its relationships to observable variables, i.e. they implicitly adopt a classical approach to measurement. However, also as noted above, because of the subjectivity involved in the choice of manifest variables, and of the form of model relating the latent and manifest variables, we might prefer to regard such statistical tools as yielding an operational definition.

According to the classical theory measurements are always real numbers: if we have been able to measure them, the numbers which have resulted satisfy all the properties required for arithmetic manipulation, so that we can manipulate them by using any statistical operation. This is as true for latent variable scores—measures of a hypothesized underlying quantitative attribute—as it is for straightforward observables such as length or weight. It is also true for measures such as preference scores—they are held to be measurements of a quantitative preference attribute, though with measurement error and possible bias, which may indeed be non-linearly related to the attribute's value. Such bias, non-linearity and measurement error can be investigated by refining the theory in which the preference scale is embedded—by relating the scores to other variables—and by using subtle statistical methods.

Luce *et al.* (1990) described *index measurement* as the use of proxy variables which are understood and which are easily measurable to act as indicants of others (not to be confused with the notion of *index numbers* mentioned in the preceding section). They gave the example (Luce *et al.* (1990), p. 323) of measuring 'hunger' by using

> 'amount of food ingested, initial rate of ingestion, force exerted to overcome a restraint to reach food, percentage reduction in normal body weight, time since last food ingestion, etc.'.

Equally, though, and with as much justification, we could use log(amount of food ingested) as a measure of hunger. This has as much empirical justification as simple amount of food ingested. So, although the amount may be measured on a ratio scale, when regarded as a direct measure of (the classical, underlying, additive scale of) *hunger* it is inappropriate to regard it as a ratio scale. It is at best ordinal. Presumably a classical theorist would postulate a theory or seek further information which would permit the measured amount (ratio scale) to be linked to the underlying attribute hunger (ratio scale).

To Adams (1966) measurements were also merely indicators (good or bad) of the underlying phenomena. He started from the premise that things like the intelligence quotient IQ are measurements and then considered what sort of measurement theory justifies this. This is the opposite of the British Association approach (Ferguson *et al.*, 1940), which started from the premise that measurement theory was (a restricted subset of) representational measurement theory and then pointed out that psychological 'measurement' did not conform—and hence was not measurement. To Adams, measurements provided systematic and objective indices of phenomena (and numbers are not essential to this). This indexical nature is explicit in areas such as economics and psychological rating scales. Laws of measurement connect the phenomenon under investigation with the results of making the measurement, but these laws need not be exactly satisfied for measurement to be useful (and, indeed, may not be exactly formulable: IQ is useful, but stating how it relates to *intelligence* is impossible). It follows that we should not ask whether a measurement procedure yields a *true* measure of the quantity but, rather, 'how good an indicator is it of the

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

phenomena it is supposed to give information about?'. That looks like classical measurement to me. Weight measurements are a good indicator, whereas preference rating on a visual analogue scale may be only ordinally related to the underlying true measure (assuming that this measure is unidimensional).

Making it quite clear that his approach was not operational, Adams asserted:

> 'measurement procedures do not *define* the concept or quantity they measure in the sense that they provide logically necessary and sufficient conditions for it. The use of a specific procedure is strictly predicated on the assumption that the basic laws of measurement . . . hold.'

Superficially similar situations arise with the use of *pointer* measurements in the physical sciences, e.g. the use of the extension of a spring to measure weight. However, as Luce *et al.* (1990) pointed out, there are ratio scale representations of weight and of length and there are theories (Newton's laws and Hooke's law) connecting the two so that using 'a spring to measure weight directly is valid according to this theory'. Expressed another way, we can put markers on the scale of length of extension which correspond to the values of weight contained within the theory, but nothing equivalent can be done in the hunger example: whatever markers we put on the length scale define the extent of hunger.

Both representational and classical views are realist, and both produce mechanistic models (which is not to say that they cannot be used to produce descriptive models). However, the representational theory maps from an assumed underlying reality and chooses numbers to produce a model of the observed relationships between 'values' of the attribute. So, only those relationships observed to exist between objects (and hence between values of the attribute) are modelled in the representational measurement system. In contrast, in classical theory the numbers are a fundamental part of the reality: relationships not directly observed between objects may also appear in the numerical system. There might, for example, be indirect evidence for such relationships. In classical theory the underlying attribute is assumed to be quantitative—and relationships between objects can be described in terms of it (perhaps via simple concatenation operations or perhaps via something more subtle). Take the case of score on a scale measuring preference for one of two alternatives. Representational measurement theory will assign people to positions (and hence numbers) on the scale and will assert that only ordinality applies and can apply. In contrast, classical theory may assert that there is an underlying quantitative variable, but that the scale is but a poor measure of it (no doubt, with an unknown non-linear relationship to it). Operational theory will define this particular type of 'preference' as being the number that emerges from the exercise.

To take another example, representational theory may assign numbers so that the ratio between the two numbers assigned to different attribute values is preserved by different numerical assignments. In contrast, classical theory will assert that the numerical value of the ratio is an empirical property of the attribute, not something *assigned*. Michell (1990) developed this argument in detail.

In the classical approach, what statistical tools we regard as appropriate to use with a particular measurement will depend on the confidence we have that the measurement procedure is accurately tapping the underlying variable. If we are confident, as for example one presumably is in measuring weight by using a balance, then we shall have no hesitation in using any statistical technique. In contrast, if we

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

suspect only a weak link, then we shall be much more circumspect about the methods that we shall use and the conclusions that we shall draw.

3. STATISTICAL STATEMENTS

In representational terms, it would in general be meaningless to compare mean preference scores based on a semantic differential scale. That is not to say that the arithmetic manipulations could not be carried out, but simply that the result would have no empirical import. In contrast, however, under the operational theory, such an analysis would be perfectly legitimate and would tell us something about the attribute 'preference' as defined by the particular scale used. This section examines such proscriptions in detail, beginning with a more detailed example.

Suppose that we take a sample of 10 rocks and rank order them according to their hardness (using, say, the Moh approach of seeing which rocks scratch which). Here the empirical system being represented is merely one of order. However, suppose that we now assign the numbers 1–10 to these rocks, in order of increasing hardness. The hardness of any new rock can be 'measured' by allocating it the number of the softest rock it is softer than. Now, in representational terms, it makes no sense to compute mean hardnesses of samples of rocks—such relationships may not be invariant to ordinal transformations and only order has been preserved by our mapping procedure. However, in operational terms, using the specified 10 rocks as a reference set, it would make sense. We can draw conclusions which are replicable by other researchers and which can be used to predict the average hardness of other sets of rocks. (Of course, if subsequent work leads to the development of a hardness scale which has more restricted invariance properties, then these operational comparisons of means may be of less interest.)

3.1. *Meaningfulness*

Representational theory hinges on a homomorphism between the empirical and numerical systems. Moreover, as described above, since the homomorphism will generally not be unique, there will be alternative legitimate numerical representations of the empirical system. Transformations between these representations are the permissible transformations. More formally, a transformation θ mapping R (the real numbers) into itself is permissible if, for every homomorphism φ from A (the attribute of the objects being studied) to R, the function composition $\theta\varphi$ is also a homomorphism from A to R. Statistical computations may be performed on the results of any of the homomorphisms, and the results will have identical substantive significance. The question then arises, what if the conclusions disagree? The classical example is calculating the arithmetic mean of data measured on an ordinal scale. The class of permissible transformations for such data is that of monotonic strictly increasing transformations. As is well known, however, the order of the means of two groups can usually vary according to the transformation employed.

Stevens's (1946, 1951) suggested solution was to restrict the statistical manipulations according to the scale type: only those manipulations which were invariant to permissible transformations were legitimate.

Unfortunately, such an approach has its problems, namely of defining exactly what is meant by a statistic being invariant (after all, change inches to centimetres

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

and the *value* of a mean changes). Such issues can be overcome, but a closer examination of this invariance approach shows that it is not the statistic *per se* which causes the difficulties, but the use to which that statistic is put—the interpretation made of that statistic or the statements made about it. The distinction, and confusion, between what is required to be able to *calculate* a statistic and what is required to *interpret* it, lies at the heart of the controversy over scale types and statistics which has rumbled on throughout most of the 20th century. Recognizing this, Adams *et al.* (1965), in an important paper, shifted attention from statistics *per se* to statements made about them. A statement is defined as being *empirically meaningful* relative to a measurement scale if and only if its truth value is invariant over permissible transformations of that scale. This is the definition of 'meaningful' that I adopt in this paper.

So, for example, a statement that one mean $\bar{x}$ is greater than another $\bar{y}$, $\bar{x} > \bar{y}$, is not generally empirically meaningful for ordinal scales: the order of the means can (usually—we shall consider a special case in a moment) be inverted by a suitable ordinal transformation. It is, however, empirically meaningful for interval or ratio scales. Conversely, although the statement $\bar{x} = \bar{z}$ is empirically meaningful for interval scales, we cannot infer that therefore it is always legitimate to use means with such scales: the statement $\bar{x} + \bar{y} = \bar{z}$ is not empirically meaningful with interval scales (though restrict the class of transformations to similarity transformations and it is meaningful). Similarly, the statement $\bar{x} \times \bar{y} = \bar{z}$, all components being measured on the same variable, is not empirically meaningful, even with ratio scales.

Thus the *context* in which a statistic is used determines whether it is legitimate or not, not merely the scale type. Sometimes a statistic is said to be *appropriate* relative to a statement and a scale if and only if the statement is empirically meaningful (invariant over permissible transformations) relative to the scale. However, if a statistical statement is or is not empirically meaningful relative to some scale, then *all* statistics are or are not appropriate respectively. The condition for appropriateness applies not only to those statistics used in the statement. We shall, presumably, be most interested in statements which *do* involve the statistic in question but the condition implies that all statistics are simultaneously either appropriate or inappropriate relative to a given statement and scale. This observation begs the question of whether we should be looking at the appropriateness of statistics at all. Perhaps the notion of appropriateness of statistics is not useful, and we should simply focus on the empirical meaningfulness of the statement (relative to the scale).

Since permissible transformations are transformations to alternative equally valid representations of an underlying empirical system, they have no place in the operational theory. Consequently the notion of empirical meaningfulness is meaningless in the operational context. (And consequently it is not sensible to ask whether a statistic is appropriate or not.) However, having said that, we can define a notion of scale type for operational measurements—in terms of the transformations which preserve the truth value of the statistical statement in question. This is outlined in Section 3.3.

Invariance over permissible transformations is an attractive definition for a statement to be empirically meaningful, but does it encapsulate all that we want? Weitzenhoffer (1951) suggested that the real criterion of 'meaningfulness' should be that a relationship could be expressed in terms of the relationships of the ERS under consideration. He argued that, in principle at least, we should be able to arrive at

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

substantive conclusions merely from consideration and manipulation of the objects themselves — but that this is generally so unwieldy as to be impracticable and hence the use of numerical representations and mathematics. Adams *et al.* (1965) also said (pages 118–119):

> 'It is worth noting that the association between empirical meaningfulness and intrinsic definability outlined here suggests an alternative way of characterising empirical meaningfulness which is to an extent independent of considerations of numerical measurement and permissible transformations. That is, a formula may be described as empirically meaningful relative to a system of measurement (more generally, any precisely formulated empirical theory) just in case it expresses a relation over the objects of the theory which is intrinsic in the sense of being definable in terms of the empirical operations and observations on which the measurement theory is based. It is a matter of conjecture that this criterion of meaningfulness is in fact more fundamental than that which defines it in terms of invariance under permissible transformations.'

Although this is clearly appealing, elucidating exactly what is meant by saying that a relationship is 'definable in terms of other relationships' is not easy. (That there is a difference is illustrated by an example of Luce *et al.* (1990), section 22.5, showing that 'definability' is narrower than invariance.) The subtlety of the issue may be illustrated by the following example.

Consider a single ordinal scale and the statement $\bar{x} < \bar{y}$ for data from it. This is meaningless in terms of the definability criterion since the arithmetic mean is not defined for ordinal scales — there is no empirical operation corresponding to addition. Also, in general, it is meaningless in terms of the invariance criterion since arbitrary monotonic transformations will change its truth value. But now suppose that we find that its truth value *is* invariant for some particular data set. For such a data set it would appear to be meaningful in terms of the invariance criterion but meaningless in terms of the definability criterion. However, a closer examination shows that for such data sets the x-sample stochastically dominates the y-sample. (Informally, this can be seen as follows. Consider the subclass of monotonic increasing transformations defined by $g(z) = k_1$ for $z < Z$ and k_2 otherwise, with $k_1 < k_2$ and z ranging over the entire range of the data. Then $\bar{x} < \bar{y}$ on all such transformed scales implies $F(z) > G(z)$, where F and G are respectively the distribution functions for the classes from which the x- and y-samples are drawn. The converse, that $F(z) > G(z)$ implies $\bar{x} < \bar{y}$ for all monotonic increasing transformations, is not true.) We could adopt the premise of this example, that $\bar{x} < \bar{y}$ for all monotonic increasing transformations, as an operational definition of one group being 'less than' another group, even though the arithmetic mean is not defined in the ERS.

In what follows the invariance definition is adopted as the defining characteristic of meaningfulness, but two points are worth stressing. Firstly, given an NRS representing an ERS, any arbitrary (invariant) relationship defined on the NRS corresponds to a matching relationship on the ERS. However, general arbitrary relationships, defined by exhaustive listing of the sets of objects satisfying them, are not of interest. What are of interest are relationships defined by means of a simple mathematical formula on the NRS. (See Luce *et al.* (1990), chapter 22.)

Secondly, in statistics we are concerned with statements about *aggregates* of objects. (Of course, these aggregates may subsequently be used to make inferential

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

statements about the behaviour of individuals, but that is a different matter, as is the fact that the objects may be aggregates themselves.) In particular, our aim is to make descriptive and comparative statements about aggregates. Now, aggregates of objects are not the same as objects. The properties of aggregates and the relationships between aggregates differ from those of individual objects. For example, a group of objects can be leptokurtic, but individual objects, pairs of objects and concatenations of objects cannot. A statement about an aggregate is an operational definition of the way of combining the individual elements (as was illustrated by the comparison of means on an ordinal scale example above). Whether this operational definition may be expressible in terms of the ERS relationships between objects need not concern us—provided that the properties of the higher level objects, the aggregates, can be unambiguously defined. (See also Hand (1993b) for discussion in the more general context of the relationship between low level and high level metadata.) Presumably one will consider it more important that the higher level definitions may be expressed in terms of the lower level relationships for mechanistic models than for descriptive models.

For some situations it is easy to show analytically that all the legitimate homomorphisms will lead to the same truth value for a statement. In others it is possible to show it, conditionally on the observed data, by exploring the permissible transformations by using numerical methods. An important subclass of the latter type of problem arises with categorical ordinal data with only a few categories. In one clinical trial that I encountered, the data had four response categories (none, mild, moderate and severe) and the question was whether there was an interaction in a study in which each subject was exposed to a 2×2 cross-classification of factors. In such a situation, all monotonic increasing transformations can be explored by fixing the two most extreme categories (at 0 and 1, say) and letting the two intermediate categories range over this interval (preserving their order, of course).

We might also reasonably expect summated rating scales, which are very popular in the behavioural sciences, to reflect a stronger empirical property than mere ordinality, though not so strong as that in interval scales. Therefore we might not be willing to countenance *arbitrary* monotonic increasing transformations. One consequence of this is that a relationship between two means computed from such a scale may have invariance properties within the subclass of transformations that we are willing to consider. For example, we may find that $\bar{x} > \bar{y}$ for *all* the transformations that we consider reasonable, if not for all monotonic increasing transformations. In essence we have identified a set of data configurations lying between those for which one group does not stochastically dominate another and those for which it does. This is the sort of situation described by Abelson and Tukey (1959) when they said:

> '*the typical state of knowledge short of metric information is not rank-order information*; ordinarily, one possesses something more than rank-order information'.

3.2. *Transformations of Data*

Given a set of objects A, suppose that the attribute under study can be represented by an interval scale variable. Call a particular numerical assignment S_1 (for example, if the attribute in question is temperature, then S_1 might be temperature measured in

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

degrees centigrade). Now, since the scale is interval, the statement that $\bar{x}_1 > \bar{x}_2$ for the mean temperatures of two groups of objects has a truth value which is invariant for all permissible transformations. This means, for example, that the statement would be true when temperature was measured in degrees Fahrenheit if and only if it was true when measured in degrees centigrade.

Now, however, suppose that a mapping from A to R, also representing the $\supseteq$ relationship by $\geqslant$, can be found such that the numbers assigned to the objects in A are normally distributed. Such an assignment (S_2, say) is also invariant up to linear transformations—i.e. any other numerical assignment such that the numbers in A are normally distributed is related to the numbers in S_2 by a linear transformation. So, again, the objects in A can be represented by an interval scale variable: again the statement that $\bar{x}_1 > \bar{x}_2$ for the mean temperatures of two groups of objects has a truth value which is invariant for linear transformations: not, note, for 'permissible' transformations—there is no empirical relationship beyond ordinality being represented and which could make linear transformations the permissible set.

We thus have two numerical assignments, both of which preserve the truth value of the statement $\bar{x}_1 > \bar{x}_2$ under linear transformations. This means that anyone who adopts the same assignment procedure as that leading to S_1 (i.e. a procedure that preserves the empirical relationships between the objects) will obtain the same truth value for this statement as was obtained using S_1. Similarly, anyone who adopts the same assignment procedure as that leading to S_2 (namely assigns numbers so that they are normally distributed) will obtain the same truth value for this statement as was obtained using S_2. Unfortunately, however, there is no reason to expect the numbers assigned by the two processes to be linearly related: the statement $\bar{x}_1 > \bar{x}_2$ might be true under assignment S_1 but false under assignment S_2.

The numbers in the first assignment process were chosen so that the relationships between them represented the relationships between the objects, i.e. the NRS was homomorphic to the ERS. In contrast, in the second assignment process, the only empirical relationship (deliberately) preserved by the mapping was order. This means that an element of arbitrariness is manifest in the assignment. We chose to remove this arbitrariness by requiring the numbers to follow a normal distribution—but this, in itself, was arbitrary—we could have chosen some other distributional form. It means that, in the second process, the chosen numerical assignment has an operational component. (We shall return to this in Section 3.3 where we consider what meaning the notion of 'scale types' might have in an operational context.) Luce *et al.* (1990) also discussed these two alternative ways of assigning the numbers.

As an example of the above, consider the following. Suppose that we want to compare the effects of two diets on the weight of calves. The comparison might be based on two groups of calves, one of which has received one diet, and the other the other diet. A direct comparison of the total weights of the two groups is not sensible unless the groups have the same numbers of calves, so one normally standardizes by the numbers in each group—and uses the arithmetic means. Various tests can then be used. In this example, with the arithmetic mean being the focus of interest, it would be wrong to use median, geometric mean or some other 'average'. Similarly, it would be wrong to transform (non-linearly) the data and then to use the arithmetic mean. In particular, it would be wrong to transform the data to normality and to perform a t-test since the test would not be comparing arithmetic means on the original scale, which is the scale which maps the empirical concatenation relationship

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

to addition. That is, the question being explored, the hypothesis being tested, would not be the question to which an answer was required. (I am indebted to Michael Healy for suggesting this example to me—he attributes it to Yates, who cautioned against a logarithmic transformation on the basis that 'the farmers are not being paid by the log(kg)'.)

More generally, a relationship of the form $y = \eta \log w$, with w the weights of the calves, is not preserved under similarity transformations of w—an extra constant term appears. Generalized linear models, with the general form $g(\mu) = \beta' x$, might at first be thought to face the same problem, but the coefficients β effectively transform the right-hand side so that it is in terms of the same units as those of the left-hand side. This is what lies at the root of dimensional analysis, mentioned above.

In the calf weight example, the original data were weights. Often, however, the scale with which we work in everyday life is a transformation of the scale representing the empirical system. Great care must then be taken to ensure that the question is stated relative to the correct scale. For example, the decibel measure of the intensity of sound is logarithmically related to sound pressure, and fuel ratings for cars are given in miles per gallon but they are measured in gallons per mile (see Hand (1994) for an illustration of the confusion that this reciprocal transformation can cause).

The fact that there may be two or more alternative, non-linearly related representations of a given set of objects, both having the same scale invariance properties, has led to some confusion. Anderson (1961), for example, remarked that 'possession of an interval scale does not guarantee invariance of interval scale statistics'. This was discussed in a stimulating paper by Velleman and Wilkinson (1993a) and further in Hand (1993a) and Velleman and Wilkinson (1993b). As pointed out in Hand (1993a), Anderson's (1961) example of whether to measure duration or speed, of which Velleman and Wilkinson said 'both are valid interval scales, and yet statistics computed on one form may be quite different from those computed on the other', tells only part of the story. If we are interested in speed then numbers preserving the relationships between speeds can be chosen and these will preserve the interval scale structure of speed. The same is true for duration if we are interested in duration. However, speed and duration are not the same things—the empirical relational structures are different, with different internal relations—so it is no surprise that analyses of the two corresponding NRSs may yield different conclusions. I pointed out in Section 2.1 that care must be taken to ensure that different alternative representations are describing the same empirical operation.

Hand (1993a) went on to say that if the researcher expects them to yield identical conclusions it suggests that the researcher believes that the two variables, speed and duration, are tapping the same underlying attribute, and it is this which is the real object of study, and not either of speed or duration. If this belief is correct, then the numerical assignments can only be ordinally related to the empirical system (they are only ordinally related to one another). Put in another way, the representation is representing only the ordering of the objects, and nothing more, so that only statements which are invariant to ordinal transformations should be made. If statements requiring interval properties are made then one can conclude, for example, that one group takes longer to reach the objective than another despite travelling more quickly. (See, for example, Hand (1994), section 3.3.) Such a contradiction is bad enough, but the situation is worse if only one of the analyses (speed or duration) is

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

carried out — then we are not even aware of the ambiguities in the conclusion. We must not fix the statistic while changing (transforming) the data without being aware that this changes the statistical model and questions, in just the same way that fixing the data and changing the statistic changes the model and questions.

This does not mean, as Velleman and Wilkinson (1993b) suggested, that I would prevent the researcher from experimenting (so that 'as a good scientist he is willing to entertain the possibility that what he thinks might not be the way the world really is') by transforming the data (by analysing both duration and speed). *Experimenting*, in this usage, means hypothesis generation, whereas *comparing*, whichever of duration or speed is used, is hypothesis testing (in a general sense). This distinction is discussed in more detail in the next section.

Generalized linear models (McCullagh and Nelder, 1989) represent an advance in this context. They establish a relationship between the conditional mean and a linear function of the covariates: the response variable is not transformed, and so remains on the original scale (representing the ERS, if that is what it does).

3.3. *Model Generation and Model Evaluation*

Evaluating a model is completely separate from formulating it. In particular, we can test meaningless hypotheses: a comparison of two means on numbers used to indicate the levels of a nominal scale, for example. Translating a scientific statement into formal statistical terms includes representing the statement numerically — in representational terms, choosing one of the homomorphisms, and, in operational terms, choosing a measuring instrument — and if the probabilistic and statistical conditions are satisfied then the statistical test is valid. It is merely that it is a test of a particular numerical assignment — and different, equally valid, assignments may lead to different conclusions. This, of course, is where distribution-free statistics come in. The hypotheses associated with distribution-free tests are typically merely in terms of ordinal relationships, so they are invariant over monotonic increasing transformations. This is worth emphasizing: the reason that distribution-free methods are appropriate for ordinal data whereas 'parametric methods' typically are not is that the former test hypotheses which can be meaningfully stated for ordinal data (are invariant to permissible transformations) whereas the latter do not. The issue is whether or not the hypotheses being tested are meaningful.

Since the operational theory does not involve notions of homomorphisms between an underlying structure and alternative legitimate numerical representations, such problems do not arise. The numbers obtained are not partly the result of the arbitrary selection of a particular homomorphism, so invariance over those that might have been selected does not apply. They are *the* numbers which emerged from the measuring instrument. Consequently the numbers which have resulted from the measurement procedure can be treated as numbers and manipulated as one will — we do not have to be constantly checking that what we are doing satisfies other constraints. The conclusions that we arrive at refer simply to those statistics calculated on numbers obtained by the specified procedure. If the numbers are subjected to a transformation, then, in effect, a different measuring procedure has been employed — the transformation is a part of the measurement procedure and a part of the definition of the concept concerned. It is no wonder, then, that the results may differ. This also illustrates why proponents of the representational and operationalist theories may reach contrary conclusions.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

In the preceding section we noted the distinction between *model evaluation* and *model generation*. In the latter, anything goes in the search for potential patterns in the data, whereas the former may be much more restrictive in what it is sensible to do—it depends on the precision with which the hypothesis is specified. The calf weight example showed that, if we wanted to test a hypothesis about arithmetic means, then it would be inappropriate to transform the data first (to normality, for example) and then to study the arithmetic means. Such a transformation would mean that the means we would be working with were not the arithmetic means of the raw data—were not the subject of the hypothesis. Implicit in the statement that we want to test arithmetic means is the fact that (in representational theory) an interval or ratio scale is involved or that an operational approach has been adopted. However, if we simply wanted to compare the 'averages' of the groups, without having a clearly specified empirical hypothesis, then arbitrary (monotonic) transformations *are* legitimate. This hinges on lack of precision in the empirical hypothesis, requiring an operational stage in defining the summary statistic, whichever measurement school we are working within. Further examples of problems arising from ambiguity in hypotheses were given by Hand (1994).

In contrast, in hypothesis generation, if we were simply seeking patterns involving summary statistics, then it would be fine to transform and use means. Any patterns discovered would relate, not to the arithmetic means of the raw data, but to some other summary statistics (e.g. to the geometric mean if a log-transformation had been used). Nevertheless, they would be patterns.

It is presumably points of this kind (imprecision in the substantive hypothesis and the arbitrariness of patterns in hypothesis generation) which led to Velleman and Wilkinson's (1993a) statement that (p. 68)

> 'Experience has shown in a wide range of situations that the application of proscribed statistics to data can yield results that are scientifically meaningful, useful in making decisions, and valuable as a basis for further research',

and why they later said (p. 70)

> 'Good data analysis . . . is a general search for patterns in data that is open to discovering unanticipated relationships. Such analyses are, of course, impossible if the data are asserted to have a scale type that forbids even considering some patterns. . . . A scientist must be open to *any* interesting pattern.'

This is all very well but for the following.

(a) Data analysis is more than simply 'searching for patterns' (and anomalies). It also involves exploring substantive models and hypotheses—we are back to the mechanistic *versus* descriptive model distinction: in mechanistic models fairly stringent constraints may be imposed on what transformations are reasonable.
(b) If the pattern exists for only an arbitrarily chosen numerical assignment then it is likely to be of no empirical interest: give me data on an ordinal scale and I can find some fascinating patterns between arithmetic means of subgroups!

Presumably Velleman and Wilkinson's point is that such patterns *may* have empirical import. However, simply exploring arbitrary transformations of the data in the hope that they will throw up 'interesting patterns' is not the most efficient way to proceed!

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

Velleman and Wilkinson (1993a) also stated (p. 68): 'There is no reason to believe that data come to us measured in the "best" way'. However, if the data have been chosen to reflect a particular empirical relational structure then, for representing that structure, they *have* been measured in the best way. Transforming a representational measurement (by a non-permissible transformation) will mean that the transformed data no longer model the ERS (at least, not by the same numerical operation). Of course, for representational measurement, if we subsequently decided that we did not wish to pose questions about the particular empirical structure being modelled, then the data might indeed not have been measured in the 'best' way. From the operational perspective, transforming the measurements will mean that the numbers are no longer on the same (perhaps the commonly used 'conventional') scale, so that statistical statements will not be directly comparable with those on raw data derived by other researchers. Best here is not defined relative to some reality, but in terms of predictive power. Thus, for operational measurements, I agree with Velleman and Wilkinson. In both situations the transformation may yield model simplification (e.g. by removing an interaction). However, if the data are regarded as having been obtained by a representational procedure the transformed numbers (after a non-permissible transformation) no longer model the internal relationships of the objects. Conversely, if the data are regarded as having been obtained by an operational procedure, the new variable is effectively defining a new measurement. In general, transformation should only be undertaken with care, and an awareness of the relationships between statements made about the transformed data and statements made about the raw data.

If we adopt a representational approach, the ERS determines the scale type of the data. This does not mean that all models and hypotheses involving those data must use all the properties of those scales. However, if the models and hypotheses make use of only a subset of the properties of the scales then they are referring only to a subset of the properties of the ERS and hence we might regard the models and hypotheses as being about a lesser ERS. It follows that we might regard the scale type as being weaker. To this extent the statistical questions might be regarded as determining the scale type. For example, we can compare medians, which use only ordinal information, of data sets measured on ratio scales but if we do this we might as well regard the data as only ordinal. This ambiguity appears to be the source of the disagreement between Velleman and Wilkinson (1993a, b) and Hand (1993a), i.e. whether we define scale type to reflect the properties of the global ERS, even though our questions may use only some of those properties, or we define scale type only in terms of the properties which are made use of in the questions. My inclination is towards the former approach, so that, for example, the fact that I use only ordinal properties in comparing the lengths of two sticks does not mean, to me, that length is only an ordinal scale.

I have two reasons for this preference.

Firstly, it discourages the inappropriate use of statistical methods for data having scale types that do not support those methods. For example, given data that are ordinal according to my definition, I will not compare two means (or, at least, not without careful thought about what we hope to achieve by such an analysis). Velleman and Wilkinson's definition would apparently allow this: after all, according to them, the ordinality of the data is not intrinsic to the data but arises only as a consequence of the questions asked of the data. This is not to say that the com-

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

parison of means on data which are only ordinal according to my definition might not throw up interesting results. But that is an issue of hypothesis generation, not of model building. By defining scale type as a property of the data (based on the way that they model reality) we can safely undertake analyses by using a subset of the ERS's properties (they will be reflected in properties of the scale type) and are protected from mistaken analyses which use properties not had by the ERS. If it is later discovered that some particular (class of) numerical assignment(s) corresponds to a stronger set of empirical relationships then naturally those assignments will be regarded as belonging to a stronger scale type.

Secondly, if we allow the scale type to be determined partly by the question then we hit unnecessary complications such as those described in section 10 of Velleman and Wilkinson (1993a). They showed that robust measures, specifically trimmed means, use interval properties of data in the central region but only ordinal properties of data in the tails. This means that, as they put it, much data would have to be described as falling into a variety of scale types simultaneously. From my perspective, however, such problems do not arise: the data have interval or ratio scale type, but for objects which lie in the tails of the distribution only a subset of the properties, namely ordinal relationships, are used.

Scale type enters the representational theory via the links between the ERS and the NRS. Since, in the operational theory, there are no such links, can scale type have any meaning in this theory? As it happens, a definition of scale type for operational measurements can be provided. To see this, let us (again) contrast the representational and operational theories.

Representational measurement starts with objects, finds relationships between them in terms of *attributes*, maps the attributes to numbers (which are values of *variables*) and makes a statistical statement about those numbers. There are constraints on what statistical statements can be made as a consequence of the requirement that the truth values of the statements must remain the same under all permissible mappings. The nature of the mappings determines the scale types.

In contrast, operational measurement starts with objects, maps the objects to numbers (which are values of variables) by using some operation (which could be the same operation as was used in the representational approach) and makes a statistical statement about those numbers. This statistical statement will have a truth value which is invariant to certain transformations and we can use that class of transformations to define the scale type of the variable. Of course, this means that the scale type will depend on the statistical statement (Velleman and Wilkinson (1993a), p. 70: 'the scale type of data may be determined in part by the questions we ask of the data'). In a sense this makes the scale type the choice of the researcher.

Let us take a real example. I have a collection of rocks. I hang them, one at a time from a spring. To each resulting extension of the spring I assign a specific number, in such a way that larger numbers correspond to greater extensions. This use of the spring provides an operational way of assigning numbers to the rocks and I shall assume that the same operation provides the numbers for both the representational and the operational measurement approaches.

For the representationalist, the rocks are the objects. The quality of causing an extension of the spring is the attribute. The number assigned is the value of the variable to be associated with that rock. The ordered nature of the numbers reflects the ordered magnitudes of the attribute. I have deliberately *not* included any attempt

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

to represent any notion of concatenation in the NRS, so any other similarly ordered sets of numbers would equally validly reflect the ordered magnitudes of the attribute. Thus the only statements which are meaningful are those which are invariant to arbitrary monotonically increasing transformations. An example of a meaningful statistical statement is that the median value of the variable for one group of rocks is greater than the median value for another group. This statement has a truth value which is invariant to any monotonically increasing transformation. The variable is thus of *ordinal* scale type.

For the operationalist, the rocks are also the objects. Numbers — the values of the variable — are assigned by noting the extension of the spring, as above. However, no other numbers have resulted or can result from this particular measurement procedure. Thus we do not have to restrict our considerations to those statistical statements which have invariant truth values under alternative legitimate numerical assignments — there are no others. So, consider an arbitrary statistical statement. Take, as an example, that the arithmetic mean of the measured variable for the rocks in one group is larger than the mean for another group. Now, this statistical statement has a truth value which is invariant to linear transformations (but not to arbitrary monotonic increasing transformations). Any alternative set of numbers, related to the original numbers by a linear transformation, will yield the same result. Thus we might describe the variable as being of *interval* scale type. But the point is that the class of transformations leading to the same truth value is determined by the statistical statement.

Thus, in the representational approach the scale type is a result of constraints implicit in the ERS, whereas in the operational approach the scale type is a result of constraints that are implicit in the statistical statement. I can find no practical value in the notion of scale type arising from the operational approach.

4. CONCLUSION

The relationship between measurement scales and statistics has been the source of much confusion and controversy. To a large extent, the confusion can be resolved by the recognition that there are several different theories of how measurement should be interpreted, just as there are different theories about how probability should be interpreted.

The representational school assigns numbers so that the numerical relationships model empirical relationships. Scale types are defined in terms of the classes of transformations between alternative numerical representations which model the empirical relationships. Statistical statements are meaningful only to the extent that they take the same truth value under different, equally legitimate, numerical representations of the empirical relationships. In contrast, in the operational school the numbers are the product of a particular measurement operation — with no reference to 'an underlying reality'. Consequently, the notion of 'equally legitimate numerical representations' is meaningless. Any numerical operation may be carried out on such numbers. Invariance of the truth value of a particular statistical statement may then be used to identify a particular class of transformations, and this may be used as a definition of scale type, but this seems of limited value. The classical school assumes that there are underlying numbers which lie on a ratio scale and it is the scientist's job to discover those numbers.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

Adding to the confusion is the practice among statisticians of referring to two different classes of structures as 'models'. As has been pointed out (e.g. by Neyman (1939), Box and Hunter (1965), Lehmann (1990), Cox (1990) and Hand (1994, 1995)) there is a difference between models that seek to represent some empirical phenomenon (which I here call mechanistic models) and models that seek merely to describe (descriptive models), though clearly there is a grey area. To a large extent, the representational and the classical theories correspond to mechanistic models whereas the operational theory corresponds to descriptive models.

Some of the confusion in the representational theory has arisen from the complication of having two *levels* of objects. At the lower level are the empirical objects under investigation. Empirical relationships exist between these objects. At the higher level are aggregates of objects, about which statistical statements are made. Whether or not the properties of aggregates can always be expressed in terms of lower level properties, the defining of aggregate properties inevitably involves an operational component.

Finally, restrictions on statistical operations arising from scale type are more important in model fitting and hypothesis testing contexts than in model generation or hypothesis generation contexts. In the latter, in principle at least, anything is legitimate in the initial search for potentially interesting relationships.

ACKNOWLEDGEMENTS

I am grateful to John Gower, John Nelder, Clifford Lunneborg and Reinhard Niederée for their comments on an earlier draft of this paper, and to Chris Wallace for some stimulating conversations about the topics dealt with in the paper. Of course, this acknowledgement does not necessarily mean that they agree with all the statements here. I also thank the five referees for their very stimulating comments, made from very different perspectives. These comments led to a substantial revision of the paper.

REFERENCES

Abelson, R. P. and Tukey, J. W. (1959) Efficient conversion of non-metric information into metric information. *Proc. Socl Statist. Sect. Am. Statist. Ass.*, 226–230.

——— (1963) Efficient utilisation on non-numerical information in quantitative analysis: general theory and the case of simple order. *Ann. Math. Statist.*, **34**, 1347–1369.

Adams, E. W. (1966) On the nature and purpose of measurement. *Synthese*, **16**, 125–169.

Adams, E. W., Fagot, R. F. and Robinson, R. E. (1965) A theory of appropriate statistics. *Psychometrika*, **30**, 99–127.

Alper, T. M. (1984) Groups of homeomorphisms on the real line. *BSc Thesis*. Harvard University, Cambridge.

——— (1985) A note on real measurement structures of scale type $(m, m+1)$. *J. Math. Psychol.*, **29**, 73–81.

——— (1987) A classification of all order-preserving homeomorphism groups of the reals that satisfy finite uniqueness. *J. Math. Psychol.*, **31**, 135–154.

Anderson, N. H. (1961) Scales and statistics: parametric and nonparametric. *Psychol. Bull.*, **58**, 305–316.

Bailar, B. A. (1985) Quality issues in measurement. *Int. Statist. Rev.*, **53**, 123–139.

Balk, B. M. (1995) Axiomatic price index theory: a survey. *Int. Statist. Rev.*, **63**, 69–93.

Bartholomew, D. J. (1987) *Latent Variable Models and Factor Analysis*. London: Griffin.

van den Berg, G. (1991) *Choosing an Analysis Method*. Leiden: DSWO.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.
Box, G. and Hunter, W. (1965) The experimental study of physical mechanisms. *Technometrics*, **7**, 57–71.
Bridgman, P. W. (1927) *The Logic of Modern Physics*. New York: Macmillan.
Campbell, N. R. (1920) *Physics: the Elements*. Cambridge: Cambridge University Press.
Cox, D. R. (1990) Role of models in statistical analysis. *Statist. Sci.*, **5**, 169–174.
Cox, T. F. and Cox, M. A. A. (1994) *Multidimensional Scaling*. London: Chapman and Hall.
Dawes, R. M. and Smith, T. L. (1985) Attitude and opinion measurement. In *The Handbook of Social Psychology* (eds G. Lindzey and E. Aronson), 3rd edn, vol. I, pp. 509–566. New York: Random House.
Dingle, H. (1950) A theory of measurement. *Br. J. Phil. Sci.*, **1**, 5–26.
Ellis, B. (1966) *Basic Concepts of Measurement*. London: Cambridge University Press.
Falmagne, J. C. (1979) On a class of probabilistic conjoint measurement models: some diagnostic properties. *J. Math. Psychol.*, **19**, 73–88.
——(1980) A probabilistic theory of extensive measurement. *Phil. Sci.*, **47**, 277–296.
Ferguson, A., Meyers, C. S., Bartlett, R. J., Banister, H., Bartlett, F. C., Brown, W., Campbell, N. R., Craik, K. J. W., Drever, J., Guild, J., Houston, R. A., Irwin, J. O., Kaye, G. W. C., Philpott, S. J. F., Richardson, L. F., Shaxby, J. H., Smith, T., Thouless, R. H. and Tucker, W. S. (1940) Quantitative estimates of sensory events. *Rep. Br. Ass. Adv. Sci.*, **2**, 331–349.
Finney, D. J. (1977) Dimensions of statistics. *Appl. Statist.*, **26**, 285–289.
Gaito, J. (1980) Measurement scales and statistics: resurgence of an old misconception. *Psychol. Bull.*, **87**, 564–567.
Gifi, A. (1990) *Nonlinear Multivariate Analysis*. Chichester: Wiley.
Gower, J. C. and Hand, D. J. (1996) *Biplots*. London: Chapman and Hall.
Hand, D. J. (1993a) Comment on 'Nominal, ordinal, interval, and ratio typologies are misleading'. *Am. Statistn*, **47**, 314–315.
——(1993b) Data, metadata, and information. *Statist. J. UN Econ. Commn Eur.*, **10**, 143–151.
——(1994) Deconstructing statistical questions (with discussion). *J. R. Statist. Soc.* A, **157**, 317–356.
——(1995) Discussion on Model uncertainty, data mining and statistical inference (by C. Chatfield) *J. R. Statist. Soc.* A, **158**, 448.
von Helmholtz, H. (1887) Zählen und Messen Erkenntnis-theoretisch betrachtet. In *Philosophische Aufsätze Eduard Zeller Gewidmet*. Leipzig. (Engl. transl. *Counting and Measuring* (1930), by C. L. Bryan. Princeton: van Nostrand.)
Hölder, O. (1901) Die Axiome der Quantität und die Lehre vom Mass. *Ver. Verh. Kgl. Sachisis. Ges. Wiss. Leipzig Math-Phys. Cl.*, **53**, 1–64.
Holman, E. W. (1971) A note on conjoint measurement with restricted solvability. *J. Math. Psychol.*, **8**, 489–494.
Krantz, D. H. (1964) Conjoint measurement: the Luce–Tukey axiomatization and some extensions. *J. Math. Psychol.*, **1**, 248–277.
Krantz, D. H., Luce, R. D., Suppes, P. and Tversky, P. (1971) *Foundations of Measurement*, vol. 1, *Additive and Polynomial Representations*. New York: Academic Press.
Kyburg, H. (1984) *Theory and Measurement*. Cambridge: Cambridge University Press.
Lehmann, E. L. (1990) Model specification: the views of Fisher and Neyman, and later developments. *Statist. Sci.*, **5**, 160–168.
Lord, F. M. (1953) On the statistical treatment of football numbers. *Am. Psychol.*, **8**, 750–751.
Luce, R. D., Krantz, D. H., Suppes, P. and Tversky, A. (1990) *Foundations of Measurement*, vol. 3, *Representation, Axiomatization, and Invariance*. San Diego: Academic Press.
Luce, R. D. and Narens, L. (1983) Symmetry, scale types, and generalizations of classical physical measurement. *J. Math. Psychol.*, **27**, 44–85.
——(1985) Classification of concatenation measurement structures according to scale type. *J. Math. Psychol.*, **29**, 1–72.
Luce, R. D. and Tukey, J. W. (1964) Simultaneous conjoint measurement: a new type of fundamental measurement. *J. Math. Psychol.*, **1**, 1–27.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.
Michell, J. (1986) Measurement scales and statistics: a clash of paradigms. *Psychol. Bull.*, **100**, 398–407.
——(1990) *An Introduction to the Logic of Psychological Measurement*. Hillsdale: Erlbaum.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

Mosteller, F. and Tukey, J. W. (1977) *Data Analysis and Regression*. Boston: Addison-Wesley.
Narens, L. (1974) Measurement without Archimedean axioms. *Phil. Sci.*, **41**, 374–393.
——— (1981a) On the scales of measurement. *J. Math. Psychol.*, **24**, 249–275.
——— (1981b) A general theory of ratio scalability with remarks about the measurement-theoretic concept of meaningfulness. *Theory Decisn*, **13**, 296–322.
Narens, L. and Luce, R. D. (1986) Measurement: the theory of numerical assignments. *Psychol. Bull.*, **99**, 166–180.
Nelder, J. A. (1990) The knowledge needed to computerise the analysis and interpretation of statistical information. In *Expert Systems and Artificial Intelligence: the Need for Information about Data*. London: Library Association.
Neyman J. (1939) On a new class of 'contagious' distributions, applicable in entomology and bacteriology. *Ann. Math. Statist.*, **10**, 35–57.
Niederée, R. (1994) There is more to measurement than just measurement: measurement theory, symmetry, and substantive theorizing. *J. Math. Psychol.*, **38**, 527–594.
Pfanzagl, J. (1959) Die axiomatischen Grundlagen einer allgemeinen Theorie des Messens. *Schrift. Statist. Inst. Univ. Wien*, **1**.
Rasch, G. (1977) On specific objectivity: an attempt at formalizing the request for generality and validity of scientific statements. *Dan. Yrbk Phil.*, **14**, 58–94.
Roberts, F. S. and Luce, R. D. (1968) Axiomatic thermodynamics and extensive measurement. *Synthese*, **18**, 311–326.
Stevens, S. S. (1946) On the theory of scales of measurement. *Science*, **103**, 677–680.
——— (1951) Mathematics, measurement, and psychophysics. In *Handbook of Experimental Psychology* (ed. S. S. Stevens). New York: Wiley.
Stine, W. W. (1989) Meaningful inference: the role of measurement in statistics. *Psychol. Bull.*, **105**, 147–155.
Suppes, P. (1951) A set of independent axioms for extensive quantities. *Port. Math.*, **10**, 163–172.
Suppes, P., Krantz, D. H., Luce, R. D. and Tversky, A. (1989) *Foundations of Measurement*, vol. 2, *Geometrical, Threshold, and Probabilistic Representations*. San Diego: Academic Press.
Suppes, P. and Zinnes, J. L. (1963) Basic measurement theory. In *Handbook of Mathematical Psychology* (eds R. D. Luce, R. R. Bush and E. Galanter), vol. 1, pp. 1–76. New York: Wiley.
Townsend, J. T. and Ashby, F. G. (1984) Measurement scales and statistics: the misconception misconceived. *Psychol. Bull.*, **96**, 394–401.
Velleman, P. F. and Wilkinson, L. (1993a) Nominal, ordinal, interval, and ratio scales typologies are misleading. *Am. Statistn*, **47**, 65–72.
——— (1993b) Reply to comments on Velleman and Wilkinson. *Am. Statistn*, **47**, 315–316.
van der Ven, A. H. G. S. (1980) *Introduction to Scaling*. Chichester: Wiley.
Von Neumann, J. and Morgenstern, O. (1947) *Theory of Games and Economic Behaviour*. Princeton: Princeton University Press.
Weitzenhoffer, A. M. (1951) Mathematical structures and psychological measurements. *Psychometrika*, **16**, 387–406.
Wise, N. M. (1995) *The Values of Precision*. Princeton: Princeton University Press.

DISCUSSION OF THE PAPER BY HAND

D. J. Bartholomew (London School of Economics and Political Science): Given that measurement, in some sense, is an essential prerequisite of statistics it is surprising that it has received so little attention from statisticians. Perhaps, like many working scientists, we view debates about the philosophical bases of our subject as of little practical importance. After all, we have managed very nicely so far and probably feel that our work is so firmly anchored in the real world that the measurement question can look after itself. The measurement of probability is an obvious counter-example to that line of argument but the author has deliberately excluded that from his paper.

What, then, are the practical benefits of this work? One concerns the matter of 'permissible statistics'. Whether or not certain statistical manipulations are legitimate turns on what view we take of the measurements on which they operate. This is an issue which I leave to other discussants. Instead, I wish to concentrate on a second area to which we might look for practical benefits. This is the contribution that the approach might make to social measurement and here I must declare an interest as the author

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

of a forthcoming book on *The Statistical Approach to Social Measurement* (Bartholomew, 1996). In that book I have advocated a model-based treatment which, though it has many affinities with the classical approach described here, seems to me to be more closely in tune with statistical thinking.

Social scientific discourse makes great use of latent variables. Business confidence, quality of life and intelligence, for example, figure prominently in social theorizing yet are not susceptible to direct measurement. Does the classification described in the present paper clarify the nature of such variables and help to place them on a firmer scientific footing? Not, I think, to any significant extent. The distinction between the three approaches is most clearly seen in the physical sciences — where it is least needed! For practical purposes it then matters little which we adopt. As we move into the social sciences, where help is most needed, the boundaries become blurred. Latent variables provide a particularly good illustration of the difficulty. In so far as they are defined by their relationships with manifest variables their measurement seems to be an example of the classical approach. But since there is an inevitable arbitrariness about which manifest variables we choose for the purpose, the author opts for the operational classification. This is superficially attractive because it settles all social measurement questions by *fiat*. But could any set of rules command universal support? Many of the concepts seem to be more firmly rooted than this arbitrariness allows and to ignore it savours of desperation rather than genuine science.

As the author notes, there have been some efforts to place social measurement on a par with that used in the physical sciences and the contribution by Rasch, referred to by the author, is a particularly noteworthy attempt. In a very limited range of circumstances it is possible to justify an interval level scale without introducing concatenation. But if social measurement is to be possible at all it needs a broader foundation.

My preference is to start within the classical paradigm by defining social variables in terms of their relationships with other variables — observable or not. These relationships are expressed by a statistical model in which the quantities of interest appear either as random variables (at the individual level) or as parameters (at the population level). The model here envisaged is what the author calls a mechanistic model, describing how the system works. The process of measurement then resolves itself into one of statistical estimation or prediction. The inevitable arbitrariness in the choice of which variables to observe and in the uncertainties of estimation and prediction are then handled as problems of sampling whether of individuals or variables. A good measurement model is one which fits the data and which survives testing in many different situations. It must also pass the ultimate test of whether it adequately translates the qualitative idea into quantitative terms. This is made easier if this is done by first expressing the essence of the idea in axiomatic form as Balk (1995) did for price levels and Shorrocks (1978) did for social mobility. The reality, or otherwise, of the measure is then on a par with all other statistical entities with which we deal. Instead of speaking of 'statistics and the theory of measurement' we would then be speaking of 'the statistical theory of measurement'. For all practical purposes we could then leave arguments about classification to others.

As tradition demands I have done my best to be critical but, whatever view one takes, the paper provides an admirable framework for debate. It is good that the Society has provided the forum for this to take place and I have great pleasure in proposing the vote of thanks to the author.

M. J. R. Healy (Harpenden): To what extent does measurement theory impinge on statistical practice? One area where it should be directly relevant is that of scale construction. An example is the assessment of physical maturity in children (Tanner *et al.*, 1975). If you X-ray the wrist of a newborn baby, you will find that it has almost no bones. The bones of the wrist appear and grow as the child's age increases until they assume their adult forms during the teenage years. There are more than a dozen separate bones and each passes through a number of recognizable stages. It is natural to suppose that the bone stages reflect an underlying property which may be called maturity (Tanner, 1959) and which in a given wrist takes a value between 0% and 100%. The problem then arises how to measure maturity from a particular X-ray. Presumably we must attach a number to each stage of each bone and then combine these numbers in some way. If we follow up the suggestion that all the bones reflect a single underlying quantity, then it seems natural to take the mean, possibly weighted. In choosing the scores to be allotted to the stages of each bone, Professor Goldstein and I (Healy and Goldstein, 1976), with the same suggestion in mind, proposed minimizing the within-subject variability, totalled over a large standardizing sample. A further decision is needed, however, since zero variability is readily achieved by choosing all the scores to be equal. Goldstein and I showed that this could be avoided by imposing a constraint on the scores. One possibility is a quadratic constraint, say that the sum of squares of all the

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

mean scores be non-zero. This leads to a method that is essentially due to Guttman (Torgerson (1958), pages 338–345). However, normal individuals change as they age from being totally immature to being totally mature. We can thus impose a linear constraint that the mean scores corresponding to the two extremes should be (say) 0 and 100. The two different types of constraint lead to rather different sets of scores.

How, then, do the theories of measurement relate to this problem? It could, I suggest, be subsumed under all three of them. As I have described it, the problem is classical; a child possesses a certain amount of maturity and the problem is to measure this. Yet it could also be described as operational. There is no external definition of maturity; instead it is defined by the methodology itself. I am less clear about the relevance of the representational theory, but I suspect that this could be pursued by taking into account the correlates of maturity and the purposes for which it is being measured. In the language of clinical trials (Schwartz *et al.*, 1980; Healy, 1978) maturity is a very pragmatic concept and does not lend itself to the more explanatory approach associated with representational measurement.

Yet the distinctions do not seem to me to be very helpful in relation to the problems associated with the measurement of maturity. I have already mentioned the two possible systems of constraint, and I could add a disquiet over the assumption of a single dimension of maturity—it is biologically plausible that the round bones of the wrist and the long bones of the forearms and fingers mature at slightly different rates. When analysing maturity data, problems arise because of the ceiling at 100% and for some purposes it proves useful to treat such data as censored, indicating a state of 'supermaturity' with a true score exceeding 100. I cannot at the moment see what light measurement theory throws on all this.

I have avoided numbering the bone stages from 1 upwards. It is tempting to do this and to use these numbers as scores, giving rise to the so-called Likert index (Torgerson, 1958). This ignores the distinction between the cardinal numbers 1, 2, 3, . . . and the ordinals 1st, 2nd, 3rd, ... A problem that preoccupies much of the measurement literature is the legitimacy of doing something like this, such as comparing the means of two sets of ranks. It seems to be widely agreed that this is not a proper thing to do. Yet applied statisticians do this whenever they utilize the standard nonparametric tests—Wilcoxon, Mann–Whitney and (transparently) Spearman and Kruskal–Wallis.

Many of the emphases in Professor Hand's paper would change quite radically if he gave more prominence to estimation at the expense of significance testing. Frank Yates's warning against a particular use of a log-transform was based on the externally imposed requirement to estimate an arithmetic mean or difference of two means. In other circumstances, an analysis of the same data based on geometric means might have been equally appropriate.

I believe that closer interaction between statisticians and measurement theorists can produce benefits for both parties. We owe a debt to Professor Hand for bringing the issues before us and I am pleased to second the vote of thanks.

The vote of thanks was passed by acclamation.

J. L. Hutton (University of Newcastle): I wish to consider the implications of this paper for the theory of measuring evidence, where the measuring device might be a jury or a judge.

In operational terms, we wish to know how different presentations of evidence given by advocates on the opposing sides can be used to predict the verdict of the jury or judge. In representational terms we wish to understand why different presentations of the evidence affect jurors, to improve our success as a barrister. We wish to be able to attain a given outcome by understanding the measuring device.

The distinction between objects and the properties of aggregates of objects will be important. Indeed, a further level will be required, as jurors might have to combine representational evidence from scientists (e.g. Balding and Donnelly (1995)) and operational evidence from psychologists. I was interested to see in the first number of the journal *Expert Evidence* that scientific reports are specifically excluded.

Juries and judges also have to measure testimony. Whether accepting testimony is a valid way of gaining knowledge has been debated since Plato. The common wisdom is that sciences, and subjects aping science, rely on *measuring*. This has been challenged recently by Coady in his book on testimony (Coady, 1994). In particular, he notes that psychologists, in aiming to be scientific by basing theories on observed measurements, claim that the testimony of witnesses is unreliable but use testimony to justify this claim.

The distinction between representational and operational schools might also contribute to

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

understanding the role of induction and David Hume's well-known assertion that knowledge cannot be gained by induction. Mechanistic models use deductive logic to represent a system. Descriptive models, used for prediction, essentially use *in*duction. In terms of Hume's dismissal of the possibility that testimony can establish the occurrence of miracles, we can usefully distinguish between two different prior distributions:

(a) Pr(What one does not expect | a mechanistic model of what can happen) or
(b) Pr(What one does not expect | our belief in a 'law of averages').

As Professor Healy does, I have reservations about Professor Hand's claims about statistical statements and meaningfulness. To say that what is meaningful is related to the statistical statement depends very much on what is being thought of as a statistical statement.

Peter J. Cameron (Queen Mary and Westfield College, London): An interesting theoretical point is alluded to by Hand: why are the nominal, ratio, interval and ordinal scales and their transforms the only scales ever used?

In Section 2.1, the results of Luce, Narens and Alper (see Alper (1987)) on scale type are mentioned. If a real scale has type (k, l) (i.e. the degree of homogeneity is k and the degree of uniqueness is l), with k and l finite and positive, then $(k, l) = (1, 1)$, $(1, 2)$ or $(2, 2)$, and the scale is determined up to a transformation (with some ambiguity in the case $(1, 2)$). But what if we use a different number system for measurement? Cameron (1989) and Macpherson (1996) have shown that, if rational numbers are used, then every type (k, l) satisfying the obvious necessary condition $k \leqslant l$ actually occurs.

Why have none of these exotic scales ever been used? As Hand points out, the structure of the empirical relational system determines the scale, since measurements are homomorphisms to the numerical system. No natural examples of empirical relational systems giving rise to strange rational scales have been discovered, but no compelling reason for their non-existence has been given.

In any case, why do we use the real number system, rather than the rationals or some subsystem thereof? Hölder's (1901) axioms lead to the real numbers, but no version of completeness can be verified empirically because of lack of precision in measurement. Certainly, any item of raw data is a rational number.

Narens (personal communication) has suggested that there may be a case for imposing some logical or topological requirement on the scale, which would exclude all but the familiar scales and their transforms. One such result, using the logical concept of o-minimality, has been found by Mosley (1996).

Stephen Senn (University College London): I disagree with some conclusions of this interesting paper. The calf feeding example is related to the issue of clinical relevance in medical statistics and reminds me of a dilemma in drug development. Should we measure how well a treatment does what we should like it to do or simply what it does? I once tended to the former view but now believe the latter is usually correct. Presumably, if feed has an effect on weight, it affects the individual calves. If it produces a near constant proportionate increase per individual, then a logarithmic transformation is useful even if (*pace*, Hand quoting Healy quoting Yates) 'farmers are not paid by the log(kg)'. A future farmer wishing to know whether to use a given diet cannot do so using the mean difference unless the mean and standard deviation of his calves are the same as those in the experiment. If, however, the proportionate increase is constant, then, given this information, it is possible to work out the economic implications for any given set of calves. A similar point to Yates's applies to meta-analysis of clinical trials.

It has been claimed, on grounds of clinical relevance, that results should be reported in terms of absolute risk rather than log-odds ratios. But, if the treatment *is* additive on the log-odds scale, the answer of a meta-analysis on the absolute risk scale is simply wrong. (It is even wrong for the average patient.) However, having analysed the data on the log-odds scale, nothing precludes us from calculating what the benefit would be for a given individual at a presumed level of absolute risk. An even simpler example can be given. A union negotiating a wage increase might agree a constant increase in pounds per individual or a constant percentage. Now, it might be that to given individuals either the absolute difference in pay or the percentage difference is considered relevant. Whichever of the two is considered more relevant, the only way that an individual can work out the implications for himself, given one single piece of information, is if the union truly reports the wage increase on the basis *actually negotiated* with the employer.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

In choosing scales of measurement for effects, we must measure the causal basis on which *change* is effected. In the calf-feeding example we are not studying the effect of weight on value but diet on weight.

Oliver Keene (Glaxo Wellcome, London): Professor Hand provides an illuminating description of theories of measurement and their implications for statistical inference.

I am surprised that the paper ignores one common aspect of measurement. People commonly assess the distance between numbers in two ways. As well as taking a simple difference, a ratio is frequently calculated. Sometimes the multiplicative system predominates, particularly in medical applications.

The paper quotes an example of a farmer giving his calves two different diets and claims that the arithmetic mean is the focus of interest. My guess is that most farmers, when presented with arithmetic means for the two groups on different diets, would immediately divide the two numbers to calculate a percentage increase. Similarly, wage increases are expressed as percentage changes yet, to misquote this example, no-one is paid by the log-pound.

I am therefore interested in Professor Hand's thoughts on this pervasive duality of methods of measuring distance between numbers.

In his section on transformations, Professor Hand implicitly implies that the only objective of a transformation is to achieve a normal distribution. As he says, choosing a transformation only because it produces a normal distribution is an arbitrary process and loosens the link to the relative value of the original numbers.

However, I am concerned that Professor Hand gives short shrift to the log-transformation and, as is frequently done, merely classes it with all other transformations. It has been argued elsewhere that there are compelling reasons for according this transformation a special status (Keene, 1995), other than simply to achieve a normal distribution.

Professor Hand commends generalized linear models because the response variable remains on the original scale. Yet a frequently used link function for continuous positive data is a log-link, and estimates of effects resulting from the modelling process are therefore ratios on the original scale. Use of any link function other than the identity link implies that the effects in the model are not additive on the original scale.

In the speed–duration example, Professor Hand appears willing only to use arithmetic means or medians. However, people often refer to distances as being twice as long and speeds half as fast. Geometric means therefore provide a third option and if used allow faster cars to reach their destination quicker. The log-transformation frequently provides a simple solution to apparent paradoxes such as this of variables measured on scales inverse to each other.

Wm Wren Stine (University of New Hampshire, Durham): I am extremely pleased to haye the opportunity to discuss this delightful paper by David Hand. The point that I wish to make concerns the confusion of two issues (Stine, 1989):

(a) should statements about empirical events be meaningful and
(b) how do we decide that a statement is meaningful within a given empirical context?

I shall argue that the answer to the first issue is yes and is certainly independent of one's philosophy of measurement or science. However, the answer to the second question is very difficult and is intimately dependent on one's philosophy of science (or measurement).

Let us consider just operationalism. Michell (1986), Hand, and, of course, Bridgman (1927), chapter 1, state that the operations used to measure an event define the event for scientific purposes. But, relationships among the results of different measurement operations are discovered and formalized into laws or theories. I have no doubt that Bridgman would claim that these laws (or theories) do not change as a function of one's philosophy of science. The algebraic axioms that define a given empirical relational structure apply to relationships among measurement operations. Indeed, Bridgman (1941), p. 19, writes

> '. . . we shall assume . . ., that two bodies each at the same temperature as a third are at the same temperature as each other, etc., etc.'.

Surely Bridgman would find the statement that a room with a measured temperature of 30 °C is twice as hot as a room with a temperature of 15 °C ludicrous. Statements about empirical events should be meaningful.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

Deciding for a given empirical situation which statements will be meaningful will depend on one's philosophy of science. Given no theory, one who operationalizes some behavioural rating scale as the attribute of interest, for example, may assume interval or ratio properties whereas someone else might imagine that the ratings are related to the underlying attributes within the person by an ordinal scale. How we decide that a statement is meaningful within a given empirical context will be intimately involved with how we define our empirical context (i.e. with the philosphy of science that we believe):

> '. . ., although these philosophies may differ with respect to how the scientist views his or her scientific statements, theories, and so on, the algebra of measurement transcends these philosophies. Indeed, only by either demonstrating inconsistencies in the mathematics or discarding mathematics as a valid form of argument can the algebra be avoided' (Stine (1989), p. 151).

C. S. Wallace (Monash University, Clayton): Professor Hand is to be thanked for reminding us of the importance of measurement in scientific enquiry, and of how the nature of a measurement limits the range of its statistical treatment. I would like to have his comments on a related point not explored in his paper, namely how a measured variable acquires its representational status. Take temperature as an example. A pseudohistory of its evolution might start with a simple categorical variable: hot, cold and ordinary. Soon this might evolve to an ordering, from very cold to very hot and burning, with equivalence modelling thermal equilibrium and an ordering induced by mixing: hot water plus cold water gives warm water. The observation of thermal expansion leads to the thermometer, and then to an operational scale that is linear in the expansion of mercury between the freezing and boiling points of water. Initially this scale could claim only to be an operationally quantified form of an order scale, but empirical evidence that the expansions of other materials are also nearly linear in this scale, and that heating and cooling rates, and the temperatures reached in mixing, have simple expressions in this scale, lead to its acceptance as an interval scale. Finally, the expansion of gases and other evidence leads to its present status as a ratio scale, with treatment as an interval scale being useful in some contexts.

During this kind of evolution, how should the analyst treat the variable? What statistics are legitimate? Further, since the status of the variable is evidently itself a matter for empirical scientific study, we should be able to use statistical methods in analysing the data used in this study, which will of course include measurements of the variable in question. We must thus ask what statistics are legitimate in studying the status of the variable being measured.

Following Professor Hand's suggestion that we form models to express and exploit 'unexpected patterns', we might argue that a variable may acquire a particular representational status, even in the initial absence of a sound theoretical basis, if assuming it to have this status leads to simple and general mathematical relationships between its measured values and the values of other variables. However, it is not obvious how we should treat the variable while trying to establish the statistical significance of such apparent relationships.

A. C. Atkinson (London School of Economics and Political Science): The questions that Professor Hand raises about measurement are particularly important in the social sciences. I, however, will discuss the material on modelling.

Generalized linear modelling is mentioned positively in the paper. It is helpful to remember that Professor Nelder's categories of measurements were produced in the context of developing GLIMPSE, an intelligent front end for GLIM.

The distinction between empirical and mechanistic models is often useful, but seemingly mechanistic models may rely heavily on simplifying assumptions. For example in chemical kinetics all reactions may be assumed first order and side-reactions may be ignored. It is difficult to believe that such models differ in essence from empirical models.

Transformations of data, for example to normality, are mentioned much less positively than is GLIM. One way of building appropriate models in science is the use of dimensional analysis and I was glad to see a reference to Finney (1977). It might be for the calf weight example that length is the underlying variable, in which case $y^{1/3}$ would be the appropriate transformation. Similarly, in the Box and Cox data on survival times, the appropriate transformation is y^{-1}, so that rate has a simple representation. If a model with suitable dimensional properties is found in which the response needs transformation for statistical purposes, such as obtaining constant variance or normality, the dimensional relationships can be preserved by transforming both sides of the equation (Carroll and Ruppert, 1988). These ideas of transformation and dimensional analysis are nicely illustrated by the MINITAB tree data (Ryan *et al.*, 1985). There are 31 measurements of the volume of cherry trees,

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

together with their girth (x_1) and height (x_2). Table 1 gives the residual sum of squares of the normalized variables, called $z(\lambda)$ by Box and Cox (1964), the normalization ensuring that the residual sums of squares are directly comparable.

Straightforward regression analysis indicates that model 1 is unsatisfactory and that a term in x_1^2 should be included. But residual plots show that y in model 2 should be transformed. The value of $\frac{1}{3}$ is included in the 95% confidence interval for the transformation parameter λ, giving model 4 in which both sides have the dimension of length. An alternative approach is to realize that tree trunks are shaped like cones, yielding model 3, in which both sides have the dimension of volume. Although dimensionally satisfactory the y in this model also requires transformation. The both sides technique leads to model 5. This model has only one parameter and also the smallest residual sum of squares of those in the table. The details are in Atkinson (1985) and Atkinson (1994).

K. Rennolls (University of Greenwich, London): The paper presents the representational and operational theories of measurement as alternatives and reasonably advocates a preference for restricting statistical statements to those appropriate to the natural scale of an empirical relational system (ERS), as a means of avoiding 'inappropriate' data analysis. This presupposes the existence of a satisfactory ERS.

It seems to me that theories of measurement and the meaningfulness of statistical statements cannot be properly considered apart from consideration of the epistemology and interpretation of scientific theories, e.g. ERSs. A statistician, and more generally a scientist, may take an eclectic view, accepting that there might be an ERS, but that knowledge of the ERS is not sufficiently firm to constrain the modelling process. Our perceptions of an ERS can only be formed through observation of measurements. Models of these measurements, including transformations, are the means of elucidating constructs and structure that are relevant to the ERS. The world of mathematical models is larger than that of empirical reality and statistical scientists will put themselves in strait-jackets if they restrict their data analysis to conform to preconceptions of the structure of the ERS.

A parallel debate to that on measurement theory in psychometrics has taken place since the 1920s over meaning, interpretation and measurement in quantum mechanics. Schroedinger's equation may be cast in the role of an ERS, whereas the Heisenberg approach was 'operational' in more than one sense. Epistemological questions have been important; does the *complex* wavefunction have an empirical (physical) existence? However, the equivalence of the approaches means that there is no basis for preference of either approach. Similarly, there will be equivalences between statistical models formulated in the natural scale of an ERS, and apparently different 'operational' models of transformed data.

In forestry, the ERS is based on the 'aggregate object', the 'forest stand', with the primary attribute of 'top height', defined as the mean height of the 100 greatest girth trees per hectare. This definition is not consistent (invariant) under proportional changes in the unit of area and number of trees in the definition. A consistent definition might be possible in terms of a conjectured multivariate-marked point process defined on the infinite plane. Does an attribute so defined constitute part of a useful and illuminating view of the forest stand as an ERS? It would certainly present us with problems in communicating with the forester. An operational approach is preferable in practical terms, but adopting an operational approach does not stop us from believing that forest stands empirically exist and have relationships.

TABLE 1

Residual sum of squares (RSS) of normalized transformations for five models for the MINITAB tree data, showing the importance of dimensional considerations

Model	*Type*	*Response*	*Carriers*				*RSS{z(λ)}*
1	Regression	y	1	x_1	x_2		421.9
2	Regression	y	1	x_1	x_1^2	x_2	186.0
3	Cone	y		$x_1^2x_2$			180.8
4	Transform y	$y^{1/3}$	1	x_1	x_2		135.8
5	Both sides	$\log y$		$\log(x_1^2x_2)$			130.6

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

The following contributions were received in writing after the meeting.

George A. Barnard (Colchester): Those of us who have come to statistics from the natural sciences or engineering have learned that most of what can usefully be said about measurement in general can be found in Campbell (1920) and comments on it by Tukey and others. See, for example, Barnard (1968), especially Tukey's discussion. But we must be grateful to Professor Hand for his detailed account of the very mixed set of doubtful ideas initiated by S. S. Stevens.

The addition in the later middle-ages of digit sequences and elementary arithmetical signs to European alphabets made possible remarkable gains in our capacity to describe the world that we live in. Poets such as e. e. cummings and novelists such as James Joyce have stretched the use of letters beyond what was formerly customary and have risked failures in communication with at least some degree of success. Why should we not allow similar freedom in the use and manipulation of digit sequences? It is for authors and readers of any such use to judge the success or otherwise of any particular case, just as it is for readers of modernist literature.

Anyone criticized for violation of Stevens's rules may refer to the history of thermometry. When Celsius introduced his thermometer it was used as an interval scale, measuring the amount of heat in a body of water. On this basis the theory of heat made much useful progress. But after the invention of the steam-engine, and the determination of the mechanical equivalent of heat, it came to be realized that temperature in kelvins should be regarded as forming a ratio scale, relating degrees Celsius to kelvins by the linear relationship $X\ ^{\circ}\mathrm{C} = (X + 273.1)$ K, for temperatures between 0 °C and 100 °C.

A point often overlooked, especially by mathematical statisticians, is that any digit sequence representing an *observation* must be finite. Huxley's monkeys typing all the books in the British Museum will eventually succeed; but they will never succeed in typing the value of π even though their typewriters can print all the digits from 0 to 9. Thus, contrary to the classical theory as quoted on p. 458, seventh line, numerical observations must be thought of as small *intervals* of real numbers. So observations on an x which can take values near to 0 can never be equated to observations on $1/x$. This fact will be borne in on anyone who looks for a cheap and accurate instrument for measuring small resistances.

D. R. Cox (Nuffield College, Oxford): Professor Hand's paper is a very welcome critical account of an interesting body of work. An additional general reference is Duncan (1984), which is partly a historical review and partly a critique of S. S. Stevens's typography of variables. As Professor Hand remarks, it is surprising that the main statistical literature does not refer to the topic more commonly; a partial explanation may be the relatively philosophical tone. The various kinds of validation bear on the fundamental distinction in multivariate analysis between internal and external methods. The point that for extensive variables (called in the paper concatenated) the mean is an appropriate parameter regardless of distributional shape is of quite wide relevance. Most importantly, however, it would be valuable to have Professor Hand's comments on the implications of the work for constructing instruments. How many dimensions are necessary to capture a particular notion? How does one set about systematically studying the relative merits of three-, five-, seven-point scales and visual analogue scales? Of course there is psychometric work bearing on these matters and in some fields considerable practical experience, but some more systematic discussion could be helpful.

N. J. Cox (University of Durham): Professor Hand's report from territory explored largely by philosophically minded physicists and mathematically inclined psychologists indicates to me that statisticians have as much to contribute to this field as they have to learn from it. Systematic theory seems to lag behind knowledge based on practice. Recall that power functions $y = ax^b$ with non-integral b were awkward children for classical dimensional analysis, which seems to have put off few users over several decades (e.g. biologists studying allometric growth). Come a theory of fractional (fractal) dimensions, and they find a welcoming home.

It is 50 years since Stevens (1946) introduced the distinctions between nominal, ordinal, interval and ratio scales, leading by way of hundreds of texts and thousands of courses to many confused and inadequate notions of measurement, not only in the social and behavioural sciences, but also in several environmental sciences, including my own subject of geography. Treated in its own terms, Stevens's scheme is a four-point ordinal scale which fails to do justice to the diversity of measurement. It omits large classes of variables, including compositional data on the simplex and directional data on the circle or the sphere. It also fails to distinguish between discrete and continuous and bounded and unbounded variables.

I bear witness to puzzlement at many apparent contradictions in the literature.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

(a) Supposedly ordinal scales span an enormous range. 'Position in a race' is just one more than the number of people faster and akin to a counted (ratio) variable. A set of positions yields many observable relationships (e.g. two runners are between fourth and seventh) that would not be preserved under all increasing monotone transformations. Even adding constants seems decidedly perverse: try telling the winner that first could be relabelled ninth.
(b) Nominal scales are repeatedly said to allow only arbitrary numerical labels. Yet it is well known that assigning 0 and 1 to binary states leads to interpretable means and all manner of worthwhile analyses.
(c) Correlation for ordinal scales is a matter of monotonicity of relationship and for interval scales one of linearity. Yet Spearman rank correlation is just Pearson correlation applied to ranks, treated as though they were interval measurements.

Duncan (1984) gave a very interesting discussion and critique of the Stevens scheme.

Mark L. Davison (University of Minnesota, Minneapolis) **and Anu R. Sharma** (Search Institute, Minneapolis):

Meaningfulness and hypothesis testing with ordinal variables

As Hand states, in hypothesis testing, the issue is the meaningfulness of the null hypothesis. We define meaningfulness as follows: let X represent an observed ordinal scale and Θ represent any alternative metric related to X by a permissible transformation. If the null hypothesis is *meaningful*, the null hypotheses for X and Θ are equivalent: $H_0(X) <=> H_0(\Theta)$ for all Θ. Hence, X can be used to test the null hypothesis as expressed in any accepted metric Θ.

In a variety of disciplines, investigators have sought conditions under which various null hypotheses are meaningful. When the standard normality and equality of variance–covariance assumptions are met by the measured variables, then null hypotheses associated with the two-sample t-test, the one-way analysis of variance (ANOVA), one-way analysis of covariance and linear regression are all meaningful. A case where meaningfulness does not hold is that of the factorial ANOVA, where the standard normality and equality of variance–covariance assumptions do *not* guarantee the meaningfulness of the main effect and interaction hypotheses (Davison and Sharma, 1988, 1990, 1994; Townsend, 1990; Spencer, 1983).

Meaningfulness is related to replicability. Behavioural researchers using different measures have sometimes reached differing conclusions about factorial ANOVA effects. The null hypothesis has been retained when expressed in the metric used by one group of researchers but consistently rejected by other researchers measuring in an alternative metric (Anderson, 1974).

In our view, Hand overstates the limitations of transformations, expressing concern that results will apply to transformed variables, but not necessarily to the raw data. However, if the null hypothesis is meaningful and the transformation is permissible, then the transformed variable provides a test of the null hypothesis in every acceptable metric Θ, including the metric of the raw data.

Hand focuses on two sample means, saying that empirical meaningfulness holds only in special cases. In practice, these 'special' cases may be reasonably common. In situations where standard assumptions hold, t-tests (or one-way ANOVAs) are appropriate for ordinal variables. When standard assumptions are not met but sample sizes are large, t- and F-statistics are none-the-less robust, and researchers may examine the empirical cumulative distribution functions or the probability density functions in the two or more groups to confirm or refute the meaningfulness of any mean differences (Townsend, 1990).

John Gower (The Open University, Milton Keynes): Once again we are indebted to Professor Hand for bringing to our attention an important area which has been strangely neglected in the Society's activities. That measurement is a fundamental concern of applied statisticians is clear from the names like biometrics, psychometrics and bibliometrics. True, the measurement of uncertainty is an aspect that deserves and receives attention in all these areas but substantive measurement is at least of equal importance. As the paper makes clear, the more difficult it is to measure properties believed to be of interest, the more attention has been paid to problems of measurement; psychometricians have been particularly active in developing measurement theories. Two issues, both with extensive literatures, that arise from psychometrics are

(a) how best to combine values on many variables. This arises, for example, in combining examination marks on several subjects to produce an overall mark; it also arises in the many

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

definitions of intersample distance where the usual, but surely arbitrary, device is first to normalize each variable.

(b) The original aim of multidimensional scaling in psychometrics and ordination in ecology is, as their names suggest, to derive a scale (or at least an ordering) from distance-like data. The original distances may be derived from more fundamental measurements or they may be derived directly by comparing pairs of objects (e.g. confusion matrices or paired comparisons). Thus, there are measurement problems from the outset but other measurement problems flow from these. Firstly, a single scale rarely adequately generates the given distances and we are forced to accept two-dimensional, or more, solutions. Attempts may be made to reify two directions in the space of the solution but, generally, we have a measure of a two-dimensional entity. Secondly, in non-metric forms of multidimensional scaling it is a remarkable fact that the redundant information available when ordering all pairs of distances, or even some subset of them, suffices to produce a tightly constrained metric solution. Ordinal information in the distances has been transformed into metric information.

These kinds of problem do not immediately seem to fit into the measurement theories discussed in the paper. Thus I ask what do the theories of measurement have to say about combining multivariate measurements, what about multidimensional measures and what about the relationships between ordinal and numerical measures?

J. K. Lindsey (University of Liege): The author distinguishes between mechanistic models for understanding and descriptive models for prediction. This appears to ignore descriptive and nonparametric statistics, which are generally inappropriate for these goals. Can the author relate his classification to these branches of statistics? (Ordinal data are not restricted to distribution-free methods; several excellent parametric ordinal models are available.)

In the classification, model generation, building and testing, where the last two are more closely associated, I do not understand how a model can be generated without building it, and thus I would prefer to associate the first two.

What is the relationship between transformation of measurements and the location–scale family? Only scale transformations are possible for ratio variables. Does this imply that the location family (with its arbitrary origin), including the normal distribution, is only applicable to interval variables?

The author provides extensive examples of measurements where the representation is by the (positive) real numbers. However, a real number has an infinite number of decimals that can never actually be recorded, in spite of the author's claim ('*the* numbers which emerged from the measuring instrument'). Many of the manipulations discussed depend on this. How can statisticians pass from such a theoretical model of measurement to the finite reality of the instruments actually used?

Thus, I am not convinced that measurement can be usefully discussed without taking into account precision. Suppose that some phenomenon can be represented by a continuous variable Y and is measured using some instrument with constant precision Δ throughout its range. We are interested in some probability model $f(y, \theta)\Delta$. Any non-linear transformation, such as from duration to speed, will no longer have constant precision throughout its range. It is astonishing that a field so concerned with measurement insists on ignoring this phenomenon in its modelling and inference procedures. Of course, taking this fact of life into account implies, among other things, that sufficient statistics are no longer available and that most point estimates currently used are inconsistent.

Closely related to this, the apparent contradiction between mean speed and mean duration arises because the model from which the mean *parameter* comes has not been specified. Once it has been, any transformation is taken into account by the Jacobian.

The fundamental Bayesian measurement problem is not prior probabilities but determining on what scale to 'measure' a parameter. Unlike empirical observation, no instrument tells us what parameter scale has constant precision.

Joel Michell (University of Sydney): Hand uses my distinction between *representational*, *operational* and *classical* paradigms of measurement but fails to stress their mutual incompatibility. Aside from counting frequencies, there are three different numerical practices in the social sciences: *measurement in the classical sense* (i.e. the estimation of the ratio of some quantitative attribute to a relevant unit of measurement), *numerical coding* (in which apparently non-numerical information (e.g. an ordering of objects) is represented numerically) and, what might be called, *opportunistic numeralization* (in which

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

apparently numerical data are generated by some other set of operations (e.g. rating scales)). While these three practices co-exist, the above three paradigms of measurement are mutually contradictory. The operational is the most inclusive (implying that all three practices are measurement), the representational excludes opportunistic numeralization and the classical numerical coding as well. Most social scientists prefer the operational for this reason. Stevens's (1946) issue of permissible statistics was raised within the representational paradigm because, obviously, given the different possible varieties of numerical coding (e.g. coding a mere classification of objects *versus* coding an ordering of objects), deductively valid numerical argument forms (including patterns of statistical reasoning) leading to conclusions about the *objects* (rather than to those about the numbers assigned) will be relative to the character of the information coded in roughly the way prescribed by Stevens; otherwise contradictory conclusions could just as easily be derived. The solution to Stevens's problem was provided by Suppes and Zinnes (1963) in the sense that, in the numerical coding situation, if conclusions derived about the objects coded are restricted just to those that remain true under permissible transformations of the numbers used, then contradictions can never be derived from the same set of data. These authors muddied the waters by introducing *meaningfulness*, a confusing and redundant concept, but one that Hand inexplicably persists with. Indeed, the general problem addressed by Hand is only solvable by considering it as one of *validity of inference*: whichever of the above numerical practices we engage in, provided that the conclusions arrived at follow validity from the numerical data, any statistical procedures used are permissible, whether these conclusions are about the objects involved or just about the numerical assignments. Hand's thesis that 'different interpretations of measurement can lead to different consequences for inference' (p. 446) is mistaken. Although this controversy originated from the competing paradigms, the above solution transcends the paradigm of measurement endorsed.

Ivo W. Molenaar (University of Groningen): The name of my university department is Statistics and Measurement Theory. With great pleasure I have followed Professor Hand's attempts to bring more clarity into the concepts covered by this name. Here are some additions to, rather than objections against, Hand's paper.

Fischer (1995a, b) has presented the foundations of the Rasch model, the issue of specific objectivity and the specific problems in the measurement of change. This is an area where conceptual, philosophical, mathematical and practical problems pose an obstacle for the fully satisfactory design, implementation, data analysis and interpretation of results in any empirical study. Fischer argued that the Rasch model has some unique properties for solving these problems.

A second interesting aspect is that a combined attack on the measurement problems and the problems of the subsequent analysis given the measurements is often decidedly superior to a separate study of the two phases. If body weight and body height correlate less than 1 nobody will attribute this to the use of poor scales and yardsticks, but in social research our instruments are far more suspect. The popularity of software like LISREL or EQS for so-called covariance structure analysis is well deserved, in the sense that the joint modelling of measurement error and natural variation can be very helpful in reaching correct conclusions. Although this approach was first based on factor analysis models, it is currently also used for other latent trait models.

A third point related to Hand's topic is the consideration of explicit measurement *desiderata* and quality criteria in the process of selection of either a measurement model or a specific measurement instrument. Samejima (1995) gave a list of criteria in her presentation of the acceleration model for polytomous items. Ellis and Van den Wollenberg (1993) explored monotone local homogeneity, by which no other personal characteristic than the latent trait value may systematically influence the category probabilities of an item. Hemker *et al.* (1996) studied the largest class of item response models for which total test score and latent trait value have a monotone likelihood ratio relationship. The common idea appears to be that a couple of desirable features are formulated, the class of measurement models satisfying them is identified and a favourable test or scale is then sought within such a class.

Reinhard Niederée (University of Kiel, Bielefeld): Hand's discussion centres on Michell's juxtaposition of 'the' operational, classical and representational approach to measurement, which I believe is too coarse to settle the controversies addressed. I shall focus here on 'the representational approach', which usually is taken to imply invariance criteria that delimit the range of admissible statistics, or more precisely the range of 'meaningful' statements involving such statistics. (For details see Niederée (1994) and Niederée and Mausfeld (1996).)

Modern representational measurement theory is founded on the concept of a homomorphic

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

representation of some empirical relational system (ERS) in a numerical relational system (NRS), emphasis being on the structural properties of the ERS expressible by (almost directly testable) 'qualitative axioms' that underlie certain numerical scales and models. A sufficiently sophisticated version of a representational account of that kind (for brevity, SR) needs to incorporate elements allegedly characteristic of an operational or a 'classical' viewpoint. The relationships constituting the ERSs need not, and often cannot, be viewed as 'observable' in a naïve strict sense (even for length: error, macroscopic objects, etc.). This introduces a theoretical, or 'latent', aspect (which need not necessarily be conceived in a standard realist fashion). Furthermore, as with the operational approach, individual scales can be referred to by extending the ERS (and NRS) accordingly.

More importantly, there is no 'logic of measurement' whatsoever that could justify traditional restrictive 'meaningfulness' criteria in terms of scale types. This includes the cases of dimensional analysis and statistical statements or hypotheses. From an SR perspective, the crucial point is not simply that 'aggregates' of objects are considered (which in Sections 3.1 and 4 is taken to imply an operational account), but rather that we are interested in a new qualitative relationship Q on (aggregates of) objects. Numerical characterizations of Q in terms of the scales based on an ERS are provably invariant with respect to admissible transformations if and only if Q preserves certain symmetries of (or is extensionally 'definable' in) that ERS. In each specific instance, such an assumption amounts to a scientific *hypothesis* about Q (or to a substantive normative stipulation), which may or may not be appropriate.

A key issue rightly stressed by Hand (see also Hand (1994)) is whether a statistical hypothesis (e.g. a comparison of means or medians) is relevant to some specific goal (e.g. some stochastic model to be tested, or a practical decision problem such as Hand's calf weight example). This is not a genuinely measurement theoretic problem, although SR analyses might sometimes prove useful in this context also. The passage from Niederée (1994), p. 568, which Hand—misleadingly—quotes in Section 2.2 belongs to a discussion of this very issue; it is not about operational approaches *per se*.

Tony O'Hagan (University of Nottingham): I am grateful to Professor Hand for a fascinating insight into a topic that, as he says, statisticians should be more aware of. I found the discussion of data transformation and the example of calf weights particularly thought provoking. If under some transformation $y_i = \phi(w_i)$ the transformed data y_i may be regarded as normally distributed, conditional presumably on unknown mean μ and variance σ^2, then that is the model for the data, whether we think of it as a model for the y_is or for the original weights w_i. If the quantity of interest is mean weight, then we wish to make inference about $E(w_i|\mu, \sigma^2)$. Unless ϕ is linear, this is a function of both μ and σ^2, and so it is clear that inference about means of y_is could be quite different.

Now I may be accused of seeing almost everything as an opportunity to score points for Bayes, but I do think that a follower of the Bayesian approach is far less susceptible to 'measurement theory errors'. One always begins by modelling, and then proceeds to make appropriate inferences. If ϕ is the log-transform, we would compute posterior distributions for

$$E(\exp y_i|\mu, \sigma^2) = \exp(\mu + \tfrac{1}{2}\sigma^2)$$

for each sample. The contrast with the less disciplined frequentist teaching is ably illustrated by Professor Hand's remark that 'various tests can then be used' on mean weights, with no reference to first constructing an appropriate model. And the frequentist statistician is not taught always to ask exactly what the object of interest should be, perhaps because inference about $\exp(\mu + \frac{1}{2}\sigma^2)$ is not so easy in a frequentist framework.

Perhaps I am being unjust. Maybe the best frequentist teaching now does emphasize the discipline of modelling and then identifying just what inference is required, but if so I find Professor Hand's warnings scarcely necessary. And have frequentists learnt, as someone whose discipline requires specifying prior information has always known, that a model expressed in terms of parameters that have no natural interpretation in the real world context is telling us loudly and clearly that these parameters cannot really be the object of interest?

P. Sprent (Wormit): I can live happily with different interpretations of probability or types of measurement provided that each gives useful information. Professor Hand has referred implicitly to the way that the information content of data depends on type, source and relation to other data. In the physical sciences such links are often clear cut. For measures of lengths of rods based on equivalence

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

classes and endwise concatenation, rigidity is a minimum physical condition for a sensible interval scale representation of length. If the rods are all made of one metal and all have the same uniform cross-sectional area, then volume, density and mass relationships imply that the length measure also provides an interval scale measure of mass. If there are minor impurities in the metal, or only small fluctuations in cross-sectional area, the length measure will still be highly correlated with a true mass measure, but if the rods are made of several different metals and have a range of cross-sectional areas we need data on densities and areas as well as length to give an informative mass measure.

If rods are made from different metals, length measurements are not temperature invariant. Length differences due to temperature changes may be trivial for some purposes; they are not if you make thermostats.

In the social sciences, the information content of well-defined measures may be virtually nil. If 12% of students fail to complete their degrees at each of several universities, does the equivalence mean anything? Each university is likely to have different entry and graduation standards, to offer different courses and to have different instruction standards and laboratory facilities, etc. However, data about these and other factors may permit covariance or more complicated adjustments to failure rates that make these a meaningful operational measure of academic performance or resource wastage.

Various scores or measurements are often combined with little thought about information content or how they are related. An arbitrarily weighted mean of often highly correlated data, some of dubious relevance, is likely to be at best non-optimal, at worst a grossly misleading single indicator of overall performance. Yet this is how many newspapers produce educational and hospital league tables. Deciding what are relevant data for answering questions of interest is no trivial matter. We are indebted to Professor Hand for an excellent paper that provides a framework for better understanding and handling of data.

Elena Stanghellini (The Open University, Milton Keynes): First of all I would like to express my gratitude for this clarifying paper on a difficult and controversial subject. My question is about latent variables. I would like clarification about the extent to which latent variables are 'operationally defined by their relationships with the observed variables'. Though I agree that the main use of latent variables in the literature is in terms of constructs of their constituent components, I do not see any theoretical reasons why this should be the only possibility. I believe that the theory of global identifiability of a latent variable model leads us to say that latent variables should follow the same measurement rules as their observed counterparts.

I shall use the single-factor model as an example. Suppose that four variables are given, $Y=(Y_1, Y_2, Y_3)'$ and X, all representing the numerical representation of an empirical relational system on the basis of a given homomorphism. Suppose that only similarity transformations are permissible and that the model $Y=\beta X+\epsilon$ represents relationships between attributes of objects, with the covariance of ϵ a diagonal matrix. If X is not observed, the theory of global identifiability of a single-factor model says that the value of β can be determined as a function of the covariances between the Y-variables, provided that the mean and variance of X are given. It seems to me that the arbitrary judgment in defining the mean and the variance of X is nothing more than the arbitrary judgment in fixing the origin and the scale of any of the other observed variables in the model.

J. P. Sutcliffe (University of Sydney): As Professor Hand writes, 'The relationship between measurement scales and statistics has been the source of much confusion and controversy'. His conclusion from his comprehensive review of the literature is that

> 'To a large extent, the confusion can be resolved by the recognition that there are several different theories of how measurement should be interpreted, just as there are different theories about how probability should be interpreted'.

That is unconvincing, however, because there are problems of interpretation of statistics within the context of any one theory of measurement. The relevant issues cannot be adequately clarified except via commitment

(a) to a tenable theory of *measurement* and
(b) to an explicit conception of *statistics*.

Only then can we

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

(c) definitively explicate the bearing—legitimacy of inference—of the former on the latter.

Professor Hand's account is deficient

(a) in its eclectic tolerance of more than one contrary theory of measurement—whereas all may be false, if one is true, certainly not all can be true,
(b) in its omission of discussion of what constitutes a statistical proposition—as distinct from a mathematical or substantial empirical scientific proposition—and
(c) in its failure to develop any one instructive case in its individual detail.

As the data for statistical analysis are very often the result of the taking of measurements, and as any one of many different modes of statistical analysis may be called for depending on the nature of the investigation which produced the data in question, there may be an indefinite number of different individual questions of 'measurement and statistics' to be decided. We cannot claim *a priori* that all such cases will submit to a single mode of resolution. Accordingly, to make a start with the clarification of basic issues, detailed critical examination should be made of at least the following three test cases: descriptive statistics—e.g. *calculating the mean* of a set of values of a (proven to be) quantitative variable; numerical coding—e.g. *appraising the co-relation* of two (empirical) variables for each of which satisfaction of no more than simple order conditions—reflexivity, antisymmetry, transitivity and connectedness—can be assumed; indicants—e.g. *inferring the significance of sampled mean differences* with respect to an unobserved variable Y from analysis of values of an observed (proven to be) quantitative variable X postulated to be monotonically increasing with Y.

Paul F. Velleman (Cornell University, Ithaca): I congratulate Professor Hand on a stimulating paper. Unfortunately, he misunderstands Velleman and Wilkinson (1993). Contrary to his characterizations, we advocate neither ignoring scale type nor the haphazard use of transformations to search for significance. We do ask researchers to take responsibility for analyses made without superfluous prior assumptions.

Hand finds this too daring. He wants us to begin each data analysis by assuming that we

(a) know what attributes to measure,
(b) have assigned numbers that preserve the salient features and relationships of these attributes and
(c) know what questions to ask about these attributes.

Although we can derive axiomatic results from such assumptions, the security that they imply is illusory because the assumptions are usually false. Often the attributes measured are proxies for other attributes, the measurement instrument is not calibrated as believed, or new questions arise.

Those who think that statistics is for answering well-framed questions about well-measured, well-understood attributes of objects that appropriately represent some homogeneous population may accept Hand's *a priori* proscriptions. Those who think it possible that we did not measure the best attribute, did not measure it in the best way, or did not pose the best question—or that the population was not homogeneous, so that no simple 'attribute' was available for measurement—may prefer the freedom to do something 'impermissible' and to judge *a posteriori* whether it was warranted.

Good data analysis requires an open mind. In John Tukey's phrase, we must be willing to look in the data 'for those things we believe are not there'. But Hand wants us 'protected from mistaken analyses which use properties not had by the ERS'. 'Great care must be taken to ensure that the question is stated relative to the correct scale', he warns.

By contrast, I admit that I may not understand everything about the data. *After* an analysis, I defend both the methods and the ability of the data to support them. We do not 'allow the scale type to be determined . . . by the question', but we may *discover* new aspects of the scale; Hand will not even let us look!

Hand may analyse only attributes that he understands *a priori* and may constrain his analyses by his assumptions about measurements. But why ask others to wear that strait-jacket? We should instead champion the freedom to learn from the data, with the *caveat* that statistical analyses also examine whether the data can support the treatment that they have received.

Leland Wilkinson (SPSS Inc., and Northwestern University, Chicago): Professor Hand does not indulge in *ad hominem*, but I find it necessary to mention personal background for fear that readers of

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

his excellent survey who have not seen Velleman and Wilkinson (1993) might conclude that Paul Velleman and I are operationalists who seek meaning in random patterns. It may suit him to think of us this way, but it leads to interpretations which were not present in our paper.

I learned psychometrics in the 1970s under Robert Abelson. The measurement writings of Stevens, Luce, Suppes, Zinnes and others to which Professor Hand refers (as well as those of Tukey, Guttman, Coombs, Tversky, Estes and others to which he does not) were part of our basic training. Heretics are those who pursue orthodoxy too vigorously, however. Paul and I *assumed* that Stevens's axioms lead to appropriate data analysis if the scale type is *known*. We simply believed, and still believe, that no analyst (including David Hand) has access to such knowledge for received data. If any statistician believes that typical medical, social or industrial experiments or surveys qualify for an axiomatic analysis, then I invite him or her to consult Wallsten (1976) or similar papers from this discipline for a reality test. If you do not know this literature, be prepared to find $N<10$ and scant use of inference. The kind of procedural control that is necessary for even a tentative empirical relational system (ERS) is unavailable to the majority of researchers. No amount of preliminary, client-centred statistical consulting will elicit this information. And the popular advice to those possessing 'sloppy data' (downgrade the measurement level to *ordinal* or *nominal*) can only be considered a parody of measurement theory.

Our position is not simply defensive, however. We believe that a presupposed ERS can blind a researcher to potential surprises in the data. We discuss several examples of this in our paper. Moreover, we believe that a researcher can *discover* measurement information by intelligent use of transformations and data plotting. And we think that Professor Hand acknowledges as much when he admits

> 'If it is later discovered that some particular (class of) numerical assignment(s) corresponds to a stronger set of empirical relationships then naturally those assignments will be regarded as belonging to a stronger scale type'.

We are interested less in any *post hoc* rationalizations for these discoveries than in whether they are encouraged.

Keming Yu (The Open University, Milton Keynes): I would like to congratulate Professor Hand for his insight into this interesting topic which is easy to ignore although one often faces the measurement problem. I am particularly impressed by two points.

The first is model generation and model evaluation. Take a simple linear model as an example. Suppose that (X, Y) are connected as

$$Y = a + bX + \epsilon$$

where (a, b) are unknown parameters and ϵ is the random error. Given a set of observations, we can propose many estimating methods for parametric estimation such as least squares, maximum likelihood, least median and some robust methods, and, in contrast, we can give several methods of assessing scores such as mean-square error, log-score, i.e. $-\Sigma_i \hat{Y}_i \log\{Y_i\}$ (assume that $Y>0$), and Minkowski R-error, i.e. $\Sigma_i |Y_i - \hat{Y}_i|^R$ ($\hat{Y}$ is the predictor of Y). However, we are accustomed to using least squares for estimation and mean-square error for assessment. It is impossible to have general equivalence between the estimating methods, or general equivalence between the assessment scores. So the question is how to select the estimation method from the class of estimating methods (much work in statistics of course has been done for this so far), and the more difficult and less considered question is how to select a score from the class of assessment scores to match the estimating method.

The second point is related to the meaningfulness of measurement, and just a little to how to assess the estimating method and the reliability of a model given a measurement score. Possibly seeking a decomposition of scores and exploring both aspects of the model and estimating method based on the decomposition are one approach to this aim. Taking the example above again with mean-square error, from

$$Y - \hat{Y} = \{Y - (a + bX)\} + \{(a + bX) - \hat{Y}\}$$

we have

$$E\left\{\sum_i (Y_i - \hat{Y}_i)^2\right\} = E\left[\sum_i \{Y_i - (a + bX_i)\}^2\right] + E\left[\sum_i \{(a + bX_i) - \hat{Y}_i\}^2\right].$$

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

If the left-hand side just measures how (in)effective the predictor is on the basis of estimating the true model (accuracy), then the first term on the right-hand side measures how big the variance of the model is and thus the reasonableness of model selection that is independent of the estimating method, and the second term on the right-hand side measures how similar the estimated model is to the true model if the model is thought to be the correct model.

Bruno D. Zumbo (University of Northern British Columbia, Prince George): I applaud Professor Hand for drawing statisticians into a debate about statistics that has barely figured in the statistical literature. I was also pleased to see the relationship between the scale of measurement and statistics controversy and the controversies about how probability should be interpreted. I have for some time made that same parallel in my thinking about the problem of restricting statistics by scale type.

What makes the measurement statistics (like the Bayesian–frequentist) controversy almost impossible to resolve is that it is bathed in all sorts of assumptions about the theory and practice of science (including statistical science). These include

(a) the various perspectives on measurement (as Hand reminds us),
(b) the role (if any) of data exploration in model and theory generation,
(c) the role (if any) of hypothesis testing in the practice of science,
(d) the various theories of truth (of which correspondence theory–invariance is only one of a multitude) and
(e) whether the proper order of inquiry is to seek to answer questions about meaningfulness in terms of scale type rather than by judging a scale's type in terms of what it is meaningful on.

Unpacking the simplest of recommendations by authors in this area requires a consideration of their assumptions.

To some, once the mathematical results have been derived there is nothing to debate—the scale of measurement restricts statistical operations. Of course since the work of David Hilbert all mathematics has been axiomatic but, as Komolgorov in his *Foundations* reminds us, any axiomatic system allows for an unlimited number of concrete interpretations besides those from which it is derived. Let us not denounce, as some have, the unificationist perspective on the measurement–statistics controversy (as articulated by Michell and others) as feeble minded, naïve, undisciplined or side-stepping the issue lest one is willing to apply the same terms to the physicist who conceptualizes light as both a particle and a wave.

Finally, Professor Hand's conclusions regarding scale-type restrictions for model fitting and hypothesis testing should be tempered by results in the methodological literature that under certain *limited* conditions we can test simple hypotheses (e.g. mean differences and model fit) without scale-type restrictions (Maxwell and Delaney, 1985; Zumbo and Zimmerman, 1993; Davison and Sharma, 1988, 1990, 1994).

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

I appreciate the thoughtful and wide-ranging comments from the discussants and regret that space limitations prevent me from replying in the detail that they deserve.

Bartholomew is, of course, a leading proponent of what I would term the classical school of social measurement. In his various writings he has presented elegant ways of measuring social variables via their relationships to manifest variables. But he himself says that he prefers to '*define* social variables in terms of their relationships with other variables' (my italics). Does not the fact that he chose to use the verb 'define' suggest an element of arbitrariness, of creativity, rather than the fact that he is simply seeking to discern the value of something which already exists? It seems to me that a fundamental point is that concepts such as 'business confidence, quality of life and intelligence', to which Bartholomew refers, are not susceptible to direct measurement because they are ill defined. To measure them we need to know exactly what we are talking about. And, to do this, we need to formulate an operational definition (in terms of variables which we can measure).

Healy's methods of assigning scores to categorical attributes are those of optimal scaling—find those scores which optimize some criterion subject to certain constraints. I agree that his approach could be defined as operational. In fact I would be tempted to go further—the essential arbitrariness of the constraints make me feel uneasy about the other interpretations. By all means assert that 'a child possesses a certain amount of maturity and the problem is to measure this', but the proposed connections to the scores assigned to each individual bone seem to have a very weak theoretical basis.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

The fact that different constraints, leading to different measures, might equally be used suggests to me that each set of constraints really *defines* alternative versions of what he means by 'maturity'. His *caveat* about unidimensionality also supports this interpretation. As to the representational school, I also am less clear about its relevance in this context.

The key issue here seems to be the problem of separating 'what is something?' from 'how do you measure it?'. Operationalism overcomes this by defining the something in terms of how you measure it. The classical school seeks to define it more closely in terms of its relationships to manifest variables. The representational approach postulates that we know very clearly what it is before we set out to measure it.

I am grateful to Cameron for stressing that the restrictions on the possible scale types which can exist are based on the assumption of a mapping to the real numbers. As he, Lindsey and Barnard point out, *data* are always defined on a subset of the rationals, where these restrictions do not apply. The question is, how does this influence the statistical conclusions which may be drawn—and, if it does not, why does it not? As far as I am aware, the only approach to statistical inference which explicitly acknowledges this and attempts to handle it in a rigorous manner is the minimum message length approach of Wallace and Freeman (1987). N. J. Cox goes further and draws attention to the extension of formal systems to produce more sophisticated models and argues that Stevens's typology does not do justice to the diversity of measurement. The examples of categorizations that I presented at the start of the paper support this. Similarly, Gower, Molenaar and Sprent raise the important point about multiple measurements. I have restricted my discussion to single variables, but this is artificial. Measurements typically arise in the context of others—we are normally trying to relate different variables together. This means that scale type and constraints on what we might regard as sensible to do to the variables should really be explored for several variables simultaneously. Conjoint measurement illustrates the power of such results.

In my phrasing of the calf weight example I did not refer to the original weights of the calves. Thus I could not carry out Senn's suggestion and standardize by the original weights. If I had known the original weights, and if I expected a near constant proportionate increase per individual (per diet), then I could indeed estimate this using the mean of the log-transformed data. But this seems to be getting rather far from the question that I did consider. Similarly, in response to Keene, I was not 'claiming' that the arithmetic mean was the focus of interest, but illustrating what would happen if it was. But surely a farmer producing cattle for the beef market would be more interested in knowing that diet A yields cattle on average 50 lb heavier than those on diet B, rather than 10% heavier, with no indication of how valuable this would be in real terms. I agree that the logarithmic transformation has useful and special properties. Its role in converting two scales, related by the reciprocal transformation, to scales which are essentially equivalent can be particularly useful. (See the discussion of Hand (1994). An example is its role in psychophysiology, where there has been a controversy over whether resistance or conductance is the more appropriate measure. The distinction vanishes if the data are first log-transformed.) A valuable paper arguing the importance of the log-transform is Törnqvist *et al.* (1985).

I believe that Stine and I are in agreement—how a researcher decides that a statement is meaningful depends on his or her philosophical orientation: an operationalist might accept a particular statement as making sense in a context where a representationalist might not.

Wallace raises the important point, with which I agree, that the measurement scale of an attribute, within the representational approach, is just as much a scientific hypothesis as is any other theory. It is thus susceptible to disconfirmation and the accumulation of supporting evidence. This is particularly pertinent when we take into account the complications of measurement error.

I agree with Atkinson that mechanistic models may be unrealistic in that they may make gross assumptions. However, this does not detract from their role as (simple) models of a believed reality, as opposed to descriptive models which are solely data driven.

I am afraid that I do not agree with Rennolls's assertion that 'our perceptions of an ERS can only be formed through observations of measurements'. We can observe the empirical relationships between (concatenations of) rigid rods, without assigning numbers to those rods. I am glad that Rennolls raises the issue of quantum mechanics. The Copenhagen interpretation is very much an operational approach—and this is one reason why it is contentious.

D. R. Cox suggests that a partial explanation for the lack of discussion of the topic of measurement in the statistical literature may be its relatively philosophical tone. A similar explanation has been proposed for its lack of practical impact in psychology, where it was, at one time, expected to have a major influence. Of course, it may be that the philosophical tone reflects a philosophical reality—and

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

that the practical effect of the differences is limited. However, one might once have made similar assertions about the different interpretations of probability.

I agree with Davison and Sharma that there are circumstances in which the truth of a hypothesis on transformed data implies the truth of the corresponding hypothesis on the raw data. A classic example is comparing the raw data medians of two distributions. If these distributions can be transformed to normality with equal variances then a t-test can be used on the transformed data: equality of the means of the transformed data implies equality of the medians of the raw data.

I do not agree with Lindsey that descriptive and nonparametric statistics are inappropriate for the goals of understanding and prediction. Descriptions can be used for effective prediction and nonparametric statistics can be used for understanding and prediction.

I believe that Michell and I are in agreement: the issue is one of validity of inference. The question is whether we are concerned with making inferences about objects or about the numerical assignments. The operational approach is about the latter—and is never invalid, though it may be of limited utility. The representational approach is about the former, and by definition imposes constraints on what are sensible numerical assignments. It is these constraints which determine the notion of meaningfulness. That the different interpretations of measurement can lead to different consequences for inference follows from the fact that they are referring to different things: to the objects or to the numerical assignments themselves.

Of course, I agree with Stanghellini that, if a latent variable is defined in terms of specified manifest variables, then the permissible transformations of the former will be determined by those of the latter. But how do we decide what manifest variables to use? An operational component may enter here.

Sutcliffe's comments have made me wonder whether the word 'theory' is right. These 'theories' are not 'descriptions of the way the universe is', so that necessarily only one can be true, but are rather descriptions of the alternative ways of doing something (measurement). Thus one *assigns* numbers (representational), *defines* them (operational) or *discovers* them (classical).

I am sorry if I misunderstood Velleman and Wilkinson (1993). However, I think that Velleman goes too far in his statement about how I 'want us to begin each data analysis'. I would rather say that *if* we know what attributes to measure, and *if* we have assigned numbers preserving the salient features of the relationships, and *if* we know what questions to ask, *then* there are restrictions on what it is sensible to do with the data. Of course, the conditions may not be met—in which case relaxing the restrictions may be sensible. I am interested in Velleman's assertion that these assumptions are 'usually' false. I suspect the truth of the usually may be discipline dependent—with the assumptions often being true in the physical sciences and less often so in the social and behavioural sciences.

In turn I suspect that Velleman has misunderstood my position, which is eclectic rather than restrictive. I believe that there is a role for each of the theories of measurement. In some situations the assumptions above are justified and in other situations they are not. Hand (1994) was especially concerned with situations where the third assumption in particular was not justified. I would feel rather uneasy being represented as an out-and-out proponent of the representational position: I often assert the fact that data analysis is an art as well as a science and cannot be reduced to axiomatic mathematics.

I suspect that the difference between my position and that of Velleman and Wilkinson really hinges around how *often* each of us expects to find data which satisfy the assumptions listed by Velleman. I expect to find such data more often than they do. This difference could be put to the test. Barnard seems willing to go even further in his readiness to relax restrictions—but surely even he would accept that there are limits (one-to-one matching of a set of objects to the natural numbers, for example?).

Zumbo hits the nail on the head when he points out that an axiom system can be mapped to the world in more than one way. Probability provides the classic example for statisticians.

In answer to O'Hagan, I think that he will find that modern frequentist teaching (even that which is not 'the best') does emphasize modelling before inference. This, of itself, does not imply that there may not be more than one way of carrying out an inference—the different procedures having different properties, each of which one may or may not find attractive.

Again I would like to thank the discussants. One thing is apparent—that the debate on the relationship between measurement and statistics has not yet reached a conclusion.

REFERENCES IN THE DISCUSSION

Alper, T. M. (1987) A classification of all order-preserving homeomorphisms of the reals that satisfy finite uniqueness. *J. Math. Pyschol.*, **31**, 135–154.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

Anderson, N. H. (1974) Cross-task validation of functional measurement using judgments of total magnitude. *J. Exptl Psychol.*, **102**, 226–233.
Atkinson, A. C. (1985) *Plots, Transformations, and Regression.* Oxford: Oxford University Press.
———(1994) Transforming both sides of a tree. *Am. Statistn*, **48**, 307–313.
Balding, D. J. and Donnelly, P. (1995) Inference in forensic identification (with discussion). *J. R. Statistic. Soc.* A, **158**, 21–53.
Balk, B. M. (1995) Axiomatic price index theory: a survey. *Int. Statist. Rev.*, **63**, 69–93.
Barnard, G. A. (1968) Measurement in the social sciences. In *The Future of Statistics* (ed. D. G. Watts). New York: Academic Press.
Bartholomew, D. J. (1996) *The Statistical Approach to Social Measurement.* San Diego: Academic Press. To be published.
Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations (with discussion). *J. R. Statist. Soc.* B, **26**, 211–252.
Bridgman, P. W. (1927) *The Logic of Modern Physics.* New York: Macmillan.
———(1941) *The Nature of Thermodynamics.* Cambridge: Harvard University Press.
Cameron, P. J. (1989) Groups of order-automorphisms of the rationals with prescribed scale type. *J. Math. Psychol.*, **33**, 163–171.
Campbell, N. R. (1920) *Physics: the Elements.* Cambridge: Cambridge University Press.
Carroll, R. J. and Ruppert, D. (1988) *Transformation and Weighting in Regression.* London: Chapman and Hall.
Coady, C. A. J. (1994) *Testimony.* Oxford: Oxford University Press.
Davison, M. L. and Sharma, A. R. (1988) Parametric statistics and levels of measurement. *Psychol. Bull.*, **104**, 137–144.
———(1990) Parametric statistics and levels of measurement: factorial designs and multiple regression. *Psychol. Bull.*, **107**, 394–400.
———(1994) ANOVA and ANCOVA of pre- and post-test ordinal data. *Psychometrika*, **59**, 593–600.
Duncan, O. D. (1984) *Notes on Social Measurement: Historical and Critical.* New York: Sage.
Ellis, J. E. and Van den Wollenberg, A. L. (1993) Local homogeneity in latent trait models. *Psychometrika*, **58**, 417–429.
Finney, D. J. (1977) Dimensions of statistics. *Appl. Statist.*, **26**, 285–289.
Fischer, G. H. (1995a) Some neglected problems in IRT. *Psychometrika*, **60**, 459–487.
———(1995b) In *Rasch Models* (eds G. H. Fischer and I. W. Molenaar), ch. 2. New York: Springer.
Hand, D. J. (1994) Deconstructing statistical questions (with discussion). *J. R. Statist. Soc.* A, **157**, 317–356.
Healy, M. J. R. (1978) Is statistics a science? *J. R. Statist. Soc.* A, **141**, 385–393.
Healy, M. J. R. and Goldstein, H. (1976) An approach to the scaling of categorized attributes. *Biometrika*, **63**, 219–229.
Hemker, B. T., Sijtsma, K., Molenaar, I. W. and Junker, B. W. (1996) Polytomous IRT models and monotone likelihood ratio of the total score. *Psychometrika*, to be published.
Hölder, O. (1901) Die Axiome der Quantität und die Lehre vom Mass. *Ver. Verh. Kgl. Sachisis. Ges. Wiss. Leipzig Math.-Phys. Cl.*, **53**, 1–64.
Keene, O. N. (1995) The log transformation is special. *Statist. Med.*, **14**, 811–819.
Macpherson, H. D. (1996) Sharply multiply homogeneous permutation groups, and rational scale types. *Forum Math.*, to be published.
Maxwell, S. E. and Delaney, H. D. (1985) Measurement and statistics: an examination of construct validity. *Psychol. Bull.*, **97**, 85–93.
Michell, J. (1986) Measurement scales and statistics: a clash of paradigms. *Psychol. Bull.*, **100**, 398–407.
Mosley, A. J. (1996) Groups definable in topological structures. *PhD Thesis.* University of London, London.
Niederée, R. (1994) There is more to measurement than just measurement: measurement theory, symmetry and substantive theorizing. *J. Math. Psychol.*, **38**, 527–594.
Niederée, R. and Mausfeld, R. (1996) Das Bedeutsamkeitsproblem in der Statistik. In *Handbuch Quantitative Methoden* (eds E. Erdfelder, R. Mausfeld, T. Meiser and G. Rudinger). Weinheim: Psychologie Verlags Union. To be published.
Ryan, B. F., Joiner, B. L. and Ryan, T. A. (1985) *Minitab Handbook*, 2nd edn. Boston: Duxbury.
Samejima, F. (1995) Accceleration model in the heterogeneous case of the general graded response model. *Psychometrika*, **60**, 549–572.
Schwartz, D., Flamant, R. and Lellouch, J. (1980) *Clinical Trials.* London: Academic Press.
Shorrocks, A. F. (1978) The measurement of mobility. *Econometrica*, **46**, 1013–1024.
Spencer, B. D. (1983) Test scores as social statistics: comparing distributions. *J. Educ. Statist.*, **8**, 249–269.
Stevens, S. S. (1946) On the theory of scales of measurement. *Science*, **103**, 677–680.
Stine, W. W. (1989) Meaningful inference: the role of measurement in statistics. *Psychol. Bull.*, **105**, 147–155.
Suppes, P. and Zinnes, J. L. (1963) Basic measurement theory. In *Handbook of Mathematical Psychology* (eds R. D. Luce, R. R. Bush and E. Galanter), vol. 1, pp. 1–76. New York: Wiley.
Tanner, J. M. (1959) Boas' contributions to knowledge of human growth and form. *Am. Anthr.*, **61**, 76–111.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms

Tanner, J. M., Whitehouse, R. H., Marshall, W. A., Healy, M. J. R. and Goldstein, H. (1975) *Assessment of Skeletal Maturity and Prediction of Adult Height*. London: Academic Press.

Torgerson, W. S. (1958) *Theory and Methods of Scaling*. New York: Wiley.

Törnqvist, L., Vartia, P. and Vartia, Y. O. (1985) How should relative changes be measured? *Am. Statistn*, **39**, 43–46.

Townsend, J. T. (1990) Truth and consequences of ordinal differences in statistical distributions: toward a theory of hierarchical inference. *Psychol. Bull.*, **108**, 551–567.

Velleman, P. F. and Wilkinson, L. (1993) Nominal, ordinal, interval, and ratio scales typologies are misleading. *Am. Statistn*, **47**, 65–72.

Wallace, C. S. and Freeman, P. R. (1987) Estimation and inference by compact coding. *J. R. Statist. Soc.* B, **49**, 240–265.

Wallsten, T. S. (1975) Using conjoint-measurement models to investigate a theory about probabilistic information processing. *J. Math. Psychol.*, **14**, 144–185.

Zumbo, B. D. and Zimmerman, D. W. (1993) Is the selection of statistical methods governed by level of measurement? *Can. Psychol.*, **34**, 390–400.

This content downloaded from 132.174.254.159 on Tue, 23 Jan 2018 20:55:10 UTC
All use subject to http://about.jstor.org/terms