REMARKS ON THE THEORY OF THE MEASUREMENT AND ITS RELATION TO PSYCHOLOGY [1]

by R. DUNCAN LUCE
University of Pennsylvania (U.S.A.)

RÉSUMÉ

Remarques sur la théorie de la mesure et ses relations avec la psychologie

Une théorie de la mesure fondamentale consiste en une ordination d'entités selon l'attribut à mesurer, une structure mathématique complémentaire appliquée à l'ensemble de ces entités, un ensemble d'axiomes (ou de lois empiriques) qui relie l'ordre à la structure, une représentation numérique (isomorphisme) du système empirique, qui est construite simplement par un décompte et un calcul de limites; enfin un théorème d'unicité qui décrit les relations entre les différentes représentations de même type. Les deux exemples décrits ici en détail constituent des mesures extensives et conjointes. Tous deux impliquent l'additivité, le premier sur une opération qui combine deux entités pour en former une troisième, le second sur les coordonnées des entités. Si les deux théories s'appliquent au même ensemble d'entités, comme c'est souvent vrai en physique classique et parfois peut-être en psychologie, alors un axiome qualitatif supplémentaire mène à une relation très simple entre les deux systèmes numériques. On compare la construction d'échelles additives sur les coordonnées avec les épreuves d'absence d'interaction dans l'analyse de la variance. Entre autres choses, ceci implique une interférence entre des conditions statistiques particulières, à savoir les tests des hypothèses et l'unicité des échelles de mesure. La position que l'on souhaite adopter est que toute transformation monotone croissante — en particulier celles que donne une normalité (approchée) des distributions — puisse être appliquée aux mesures pourvu que la question posée ne soit pas relative à la nature des échelles. Par exemple : deux échantillons sont-ils tirés de la même distribution ? Toutefois, quand la question concerne l'échelle, c'est-à-dire : y a-t-il une échelle qui soit (approximativement) additive sur les coordonnées ? alors seules les transformations qui sont admissibles à l'intérieur de la théorie de la mesure peuvent être utilisées sans modifier la question. En général, ceci signifie que les hypothèses d'équinormalité de la statistique classique ne peuvent être satisfaites quand la théorie de la mesure est suffisamment forte pour fournir une échelle d'intervalle ou une échelle métrique.

[1] This work was supported in part by National Science Foundation grant NSF GB - 1462 to the University of Pennsylvania.

ABSTRACT

A theory of fundamental measurement consists of an ordering of entities according to the attribute to be measured; some further mathematical structure over the set of these entities; a set of axioms (or empirical laws) that interrelates the ordering and the structure; a numerical representation (isomorphism) of the empirical system which is constructed just by counting and computing limits; and a uniqueness theorem that describes the relations among different representations of the same type. The two examples described in some detail are extensive and conjoint measurement. Both involve additivity — the former over an operation of combining two entities to form a third, and the latter over the coordinates of the entities. When both theories apply to the same set of entities, as is often true in classical physics and occasionally may be in psychology, then an added qualitative axiom leads to a very simple relation between the two numerical systems. The construction of scales that are additive over coordinates is compared with analysis-of-variance tests for no interaction. Among other things, this involves the interplay between particular statistical statements, e.g , tests of hypotheses, and the uniqueness of the scales of measurement. The position advocated is that any monotonic increasing transformation — in particular, those that lead to (approximate) normality of distributions — may be applied to the measurements provided that the question asked is not about the nature of the scale. An example is whether two samples were drawn from the same distribution. However, when the question concerns the scale — e.g., is there a scale that is (approximately) additive over coordinates ? — then only those transformations that are admissible within the measurement theory may be used without altering the question. In general, this means that the equi-normality assumptions of classical statistics cannot be satisfied when the measurement theory is sufficiently strong to lead to interval or ratio scales.

1. Introduction

As with many basic truths, the one to the effect that science depends crucially upon our ability to measure is elusive, easily misunderstood, and continually subject to amplification and reinterpretation. And so it is not surprising that as the behavioral sciences have become more sophisticated they — and, in particular, psychology — have devoted a portion of their literature and some of their better mathematical talent to an examination of what this truth may mean for them and how to measure their own basic variables. These analyses may prove to be one of the lasting contributions of contemporary behavioral science to, at the very least, the philosophy of science. Although this work has just begun to penetrate philosophical circles — there are only a few signs of it in their journals, let alone in more discursive essays and textbooks — nonetheless it is quite clear that we understand some aspects of fundamental measurement better than did physicists, such as Campbell

(1920, 1928), or philosophers, for example Cohen and Nagel (1934), who wrote a generation ago.

Less clear is the extent to which these studies eventually may contribute directly to the behavioral sciences as such. At least two reasons underlie this uncertainty. For one, too few experimental studies have been completed for us to know how satisfactory the new theories are. For another, even if we had a theory of measurement that provided a reasonably satisfactory description of a psychological attribute, we do not have as yet suitable techniques and equipment to conduct routine measurements in the way physicists do regularly. It is rare, indeed, for a physicist to measure length or mass by direct appeal to a theory of fundamental measurement; rather, sizable portions of the theoretical superstructure of physics are used to devise alternative and vastly more convenient ways to measure these and other quantities. Practical measurement often rests upon highly refined theory and equipment whose development has, quite literally, taken decades and sometimes centuries of work. If the parallel is correct, then even were an adequate behavioral theory of measurement available — and I am not prepared to argue that any is — almost surely we do not have the superstructure of psychological theory needed to evolve a technology which could provide us with clever, rapid, and accurate means of measuring basic psychological quantities.

That being so, I can only assume that some interest attaches to our increased understanding of the nature of measurement even though it may not help much in actual everyday measurement. My comments on our understanding fall into three parts. First, just what do we mean by fundamental measurement and in what ways do our current views differ from the earlier ones ? (Secs. 2, 3, and 4). Second, what can be said about the formal structure of substantive theories, i.e., ones that establish relations among measures of several variables, when some of them are measured both physically and psychologically ? (Sec.'5). Third, what, if any, constraints do theories of measurement impose on our procedures of inference — especially, on our statistical procedures ? (Secs. 6 and 7).

2. The Nature of Fundamental Measurement

By fundamental numerical measurement I mean an assignment of numbers to objects or events and an assignment of numerical relations to qualitative relations among these objects or events such that (a) one of the qualitative relations unambiguously orders the objects or events according to the attribute one wishes to measure, and (b) the numerical relations reflect (are isomorphic to) the structure of the qualitative relations. To illustrate (a), if we wish to measure the attribute mass, then it is assumed that some unambiguous qualitative observation, such as deciding which pan of a two-pan balance drops when objects are placed on both pans, permits us to identify which of any pair of objects has the greater mass. If loudness is the attribute, then it is assumed that a subject's response as to which sound is louder can be discriminated unambiguously by the experimenter and that it identifies which of a pair of sounds is louder for that subject. As we shall see, additional qualitative relations sometimes exist among the entities to be measured, but an ordering according to the attribute of interest must always be present in a theory of measurement.

By "the structure of the qualitative relations" I mean the network of empirical laws that these relations are assumed to satisfy. In the mathematical development, such postulates are usually called "axioms" rather than "laws," but we must not lose sight of the fact that a theory of measurement possesses empirical interest only to the extent that its axioms are (approximately) true empirical laws — admittedly, of a fairly low level of abstraction. Once this structure is stated, the experimentalist and theorist follow different routes. The latter accepts the structure as axiomatic and investigates its properties, usually by constructing a numerical system that is isomorphic to the qualitative system; such a result is known as a *representation theorem.* The experimentalist, however, is concerned with the empirical adequacy of the alleged laws; this involves direct experimental tests. Examples exist where the representation theorem has been used in an attempt to evaluate a theory, but direct tests of the axioms are usually more satisfactory and are more likely to localize the defects of the theory.

To be called a theory of fundamental measurement, no numbers may enter into the empirical system; nevertheless, the representation theorem provides a numerical assignment to the elements being measured. This is possible only because we assume in the proof of the theorem that the "measurer" is able to count the number of elements in any finite set. With that implicit assumption, certain limiting processes are carried out which assign real numbers, not in general just integers or rational numbers, to the elements. Of course, in practice such an idealization breaks down for very large sets, thereby placing a bound on the precision of measurement. This is not, however, a particularly significant limitation since experimental imperfections embodied in the qualitative data usually limit the precision even more.

3. Extensive Measurement

In most satisfactory measurement systems, some additional structure is provided beyond the ordering of the attribute of interest. Classical *extensive measurement,* which is a model for mass, length, and other fundamental measurement in physics, includes an operation, known as *concatenation,* whereby two entities having the attribute to be measured can be joined together to form a third entity that also has the attribute. For example, suppose a and b are objects, the attribute is mass, P is the qualitative relation of "greater mass than" as judged by, say, which pan of an equal-arm pan balance drops, and I is the qualitative relation of "equal mass to" as judged by no movement of the pan balance. Let aob denote the new object formed by placing a on top of b. Since this also has the attribute "mass," "o" is a concatenation operation. Examples of two of the empirical laws (axioms) involving o are : (i) $(aob)\,\mathrm{I}\,(boa)$, and (ii) if $a\mathrm{P}b$, then for all c, $(aoc)\,\mathrm{P}\,(boc)$. The first (when coupled with other axioms of the system) has the interpretation that if object c exactly balances aob, i.e., "a on top of b", then it will also balance boa, i.e., "b on top of a". The second means that if the a-pan drops when a and b are placed on the balance ($a\mathrm{P}b$), then the pan in which a is placed on top of c (aoc) will drop if b is placed on top of c (boc) in the other pan, i.e., $(aoc)\,\mathrm{P}\,(boc)$. Actually, of course, b is placed on top of a c' that balances c. With a sufficient set of such axioms, all as plausible as these and as capable of (approximate) empirical verification, it can be shown (Suppes, 1951; Suppes and Zinnes, 1963; and earlier papers referenced in these two papers) that there is an assignment θ of positive numbers to objects such that, for all objects a and b,

$$\text{i.}\quad a\mathrm{P}b \text{ if and only if } \theta(a) > \theta(b),$$

and

$$\text{ii.}\quad \theta(aob) = \theta(a) + \theta(b).$$

This is the representation theorem for extensive measurement. Since an ordering relation exists in any theory of fundamental measurement and since numerical order in the numerical system is always chosen to reflect the qualitative one, Property i is a part of every representation theorem. Property ii — in this case, the additivity of mass — is, however, unique to this system of extensive measurement.

Let me outline the basic idea involved in the construction of θ. Suppose that a is any object and a_i, $i = 1, 2, ..., n$, are objects that each balance a, then let na denote an object that balances $a_1 oa_2 o \ldots oa_n$.

Choose arbitrarily some object e, and let $\theta(e) = 1$; e has, by definition, unit mass. Let a be any object. If for some integer n, $a\mathrm{I}(ne)$, then assign $\theta(a) = n$ (this must be the assignment in order for Properties i and ii to hold). If such an exact balance does not obtain, then for any positive integer n it is plausible that we should be able to find an integer m, dependent upon n, such that $[(m+1)e]\,\mathrm{P}\,(na)\,\mathrm{P}\,(me)$, i.e., $m+1$ *e's* together are heavier than n *a's*, which in turn are heavier than m *e's*. If so and if Properties i and ii are to hold, we must have

$$(m+1)\,\theta(e) > n\theta(a) > m\theta(e),$$

from which it follows immediately that,

$$\frac{m}{n} < \theta(a) < \frac{m}{n} + \frac{1}{n}.$$

Thus, to within a precision of $1/n$, $\theta(a)$ equals m/n. This estimate can be made as precise as we please by taking n as large as necessary. The proof amounts to showing that the axioms permit the various steps suggested, that the limiting process hinted at really works, and that the resulting function θ satisfies Properties i and ii.

Notice two things. Counting has played a crucial role in the proof. And the basic nature of the measurement process involves determining how many duplicates of one object balance, or approximately balance, how many of another. Both of these statements are true of all systems of fundamental measurement with which I am familiar.

The second major theorem of a theory of measurement tells us the relation between two different numerical assignments that each fulfill the conditions of the representation theorem; this is known as the *uniqueness theorem.* In extensive measurement, it is clear that we could have chosen any object other than e and called its mass 1, so multiplication of θ by a positive constant yields an equally good representation. It can also be shown that any two representations that satisfy Properties i and ii are related by multiplication by a positive constant. This is summarized by saying that mass — and indeed any extensive quantity — is measured on a ratio scale.

4. Conjoint Measurement

Because extensive measurement is the only type of fundamental measurement that has ever been proposed in physics, it was believed for a while that the two notions are synonymous. The behavioral sciences have dispelled that belief by constructing alternative schemes of fundamental measurement. The most recent, and one of some generality, is known as *conjoint measurement* (Luce & Tukey, 1964); it rests on this idea. Suppose that the alternatives that have the attribute we wish to measure can be identified by a symbol of the form (a, x), where a names one reproduceable component and x names another. For example, (a, x) might be a mass called a that is moving at a velocity called x. Neither a nor x need be numerical measures; for example, we can identify and reproduce the velocity simply by knowing the point above a fixed location from which the object was released. Or (a, x) might be a pure tone of intensity a and frequency x, where again all we need is a means to identify and reproduce intensities and frequencies, not necessarily measures of them. Let P be an ordering of these entities according to the attribute of interest, and suppose that it depends upon both components. In the first case, (a, x) P (b, y) might mean that the moving object (a, x) has greater momentum than (b, y) under some well-specified experimental conditions. Axioms about P are stated that are sufficient to prove that there are numerical functions θ over the entities, ϕ over their first coordinate, and ψ over their second coordinate such that for all (a, x) and (b, y),

$$\text{i}' \quad (a, x) \text{ P } (b, y) \text{ if and only if } \theta(a, x) > \theta(b, y),$$

and

$$\text{ii}' \quad \theta(a, x) = \phi(a) + \psi(x).$$

These scales are unique up to positive linear transformations with the same slope, i.e., they are interval scales with related units.

There is a clear family resemblance between this theory and extensive measurement, and at the same time there are important differences. The most obvious difference, and the one that makes conjoint measurement of interest to the behavioral sciences, is the fact that no concatenation operation is postulated. Its role has been assumed by the apparently much more innocent postulate that entities with the attribute to be measured have (at least) two components, each of which affects the attribute and each of which can be independently manipulated by the experimenter.

As with extensive measurement, the construction of the numerical measures involves both counting and the establishment of equivalences among the alternatives. To be specific, suppose a, a', b, and b' are from the first coordinate and that there are x and y from the second such that both

$$(a, x) \text{ I } (b, y) \text{ and } (a', x) \text{ I } (b', y)$$

hold. If so and if the conjoint representation exists, it follows readily from properties i′ and ii′ that

$$\phi(a) - \phi(b) = \psi(y) - \psi(x) = \phi(a') - \phi(b'),$$

i.e., the interval between a and b equals that between a' and b' because both equal that between x and y. In this way we can construct replicas of a given interval. If we choose some a_0 to be the zero of ϕ and some larger a_1 to be the unit of ϕ, then just as in extensive measurement, we can ask how many (adjacent) replicas of the a_0 to a_1 interval are approximately equal to how many replicas of the a_0 to b interval, and the limit of the ratio of these two integers is the value $\phi(b)$. The function ψ is constructed similarly, and one then shows that they satisfy Properties i' and ii'.

In Luce and Tukey's theory, four axioms insure that the various steps of the proof are possible; they are, roughly, the following : (*a*) that the relation R (= P combined with I) weakly orders the elements; (*b*) that R satisfies a simple cancellation property which amounts to dropping the same thing from both sides of certain pairs of inequalities; (*c*) that the elements of the two coordinates are sufficiently finely graded (or, when they are discrete, appropriately spaced) and extensive that equivalences such as (a, x) I (b, y) can be solved for the fourth element when the other three are fixed; and (*d*) that an Archimedean condition is met that says, in effect, that no finite difference, however large, is infinitely larger than any non-zero difference. If the desired representation is to hold, Axioms (*a*) and (*b*) are inescapably true, as also is Axiom (*d*) or something very much like it, since subsets of numbers exhibit the corresponding property. Axiom (*c*) is by no means a necessary condition for the representation and, what is worse, it is quite restrictive because, for all practical purposes, it requires both continuous coordinates and unbounded measures on each coordinate. However, it has recently been shown (Luce, 1966) that the same representation can be established if Axiom (*c*) is weakened considerably, another simple necessary condition (monotonicity in each component) is added, and a sufficient number of elements are postulated. The weakened form of Axiom (*c*) for the first component is : Let a, x, and y be given, then there exists a b such that (a, x) I (b, y) provided that there exist b', b'' such that (b'', y) R (a, x) R (b', y). The statement for the second component is similar [1].

[1] Other results may be briefly mentioned. When each component has only a finite number of elements, Scott (1964) and, independently, Tversky (1964) have completely eliminated Axiom (c) and have stated necessary and sufficient conditions for the existence of an additive representation. Tversky (1967) extended these results to the infinite case and to non-additive, polynomial representations. The drawback with these results is that the apparently simple conditions are, in fact, a complex bundle of cancellation conditions. Thus, from an empirical point of view, interest continues in sufficient conditions that involve only a few axioms. In addition to the sufficient conditions mentioned in the text, Debreu (1959, 1960) has given a simple system that rests in part upon topological assumptions about the components. Adams & Fagot (1959) discussed necessary conditions.

Conjoint measurement is interesting not primarily because it demonstrates that extensive measurement is a special type of fundamental measurement or because it provides physics with an alternative to extensive measurement, but rather because of its potential use in psychology and the other behavioral sciences. To my knowledge, no one has ever proposed a concatenation operation for any psychological attribute with the serious hope that the axioms of extensive measurement would be satisfied. By contrast, conjoint measurement can be tested in a variety of situations; all that is required is two variables that affect the attribute of interest. As yet, however, few attempts have been made to test it. Whether much can be done without further theoretical work is doubtful since, in general, subjects do not exhibit consistent responses, as is assumed in such an algebraic theory [1].

5. Simultaneous Extensive and Conjoint Measurement

In physics, at least, and I believe in psychology to some extent, situations exist in which some variables can be measured both extensively and conjointly; when that is so, how do the two types of measures relate ? For example, let an object of mass m and velocity v have kinetic energy w. When m and v are measured in the usual (extensive) way, it is well known that $w = mv^2/2$. Now, suppose that we have a way of deciding qualitatively which of two moving objects has the greater

Additional results about conjoint measurement that have or soon will appear are : Luce and Tukey (1964) showed that if formal "negative" entities are introduced into extensive measurement, then the resulting representation theorem is a consequence of the one for conjoint measurement; and Krantz (1964) showed roughly the converse, namely, that when the axioms of conjoint measurement are satisfied a formal concatenation operation can be introduced which satisfies the axioms of extensive measurement for positive and negative entities, and that representation provides a proof of the one for conjoint measurement. In addition, he developed a generalized symmetric theory in terms of three relations on a set of elements for which it is not necessary to identify in advance the two components. Other n-component generalizations can be found in Debreu (1960) and Luce (1966). Many of the two-component results are closely related to theorems in the algebraic theory of webs (Aczél, Pickert & Radó, 1960; also see references given there). Finally, Roskies (1965) has generalized the Luce-Tukey result to a multiplicative rather than additive representation; in general, the former cannot be reduced to the latter by taking logarithms because the scale values may be negative as well as positive.

[1] For example, Mc Laughlin & Luce (1965) (also see Luce & Suppes, 1965, for a summary of closely related empirical work on preferences) attempted to test the cancellation and transitivity axioms for preferences among bitter-sweet solutions, but the data forced them to evaluate probabilistic generalizations of both axioms. Marley (1965) has obtained some theoretical results about such probability models, but it is too early to draw any conclusions about these models. Tversky (1965) also has tested additivity within a gambling context.

kinetic energy and that the resulting order R satisfies the axioms of conjoint measurement, which it would according to classical physics, then taking exponentials of the additive representation theorem, we have measures Θ, Φ, and Ψ such that $\Theta\ (m, v) = \Theta\ (m)\ \Psi\ (v)$, where Θ is the conjoint measure of kinetic energy, Φ is the contribution of mass to this measure of kinetic energy, and Ψ is the contribution of velocity to it. So the question is : how do $\Theta\ (m, v)$ and w, $\Phi\ (m)$ and m, and $\Psi\ (v)$ and v relate ? It is clear that if they are not closely related — in fact, substantially the same — we are faced with alternative measures of the same attribute and no obvious means to choose between them.

At least for physics, the situation is fortunately fairly simple, but nonetheless interesting. In addition to assuming that R satisfies the axioms of conjoint measurement and that mass and velocity are each extensive measures, let us suppose that the two schemes of measurement are related in the following way : there are non-zero integers p and q such that for all positive integers i and j and for all m and v,

$$(i^p m, j^q v)\ \mathrm{I}\ (j^p m, i^q v), \tag{1}$$

where $i^p m$ denotes the element obtained by i^p concatenations of the entity identified by m, etc. If so and if a certain technical assumption is made, then it can be shown (Luce, 1965) that there are constants α, α_1 α_2, and $\beta > 0$ such that

$$\begin{aligned} \Theta\ (m, v) &= \alpha w^{\beta}, \\ \Phi\ (m) &= \alpha_1 m^{\beta q}, \\ \Psi\ (v) &= \alpha_2 v^{\beta p}. \end{aligned} \tag{2}$$

In this particular example, Eq. 1 holds only when $p/q = 2$ and $\alpha = 2\alpha_1\alpha_2$. Of course, the general result is not restricted just to mass, velocity, and kinetic energy and their particular constants. The free exponent β arises simply from the fact that the scales of additive conjoint measurement are interval scales with a common unit, so when the scales are exponentially transformed they are unique up to a common, positive exponent.

The conclusion is that the two theories of measurement are compatible in a very simple way provided that m and v are extensive measures and there are integers p and q such that

$$w = \alpha\ m^q v^p \tag{3}$$

induces the ordering R of conjoint measurement. If this is true, then it is easily shown that Eq. 1 is true — indeed, Eq. 1 can be interpreted as a qualitative form of the numerical law Eq. 3. Conversely, if both types of measurement theories apply, if Eq. 1 is true, and if a technical assumption holds, then the multiplicative conjoint measures relate to the extensive ones as in Eq. 2, and the numerical law, Eq. 3, is satisfied. (Luce, 1965).

This result, or ones closely related to it, may prove of value in psychology. Suppose, for example, that a subject orders — perhaps by pair comparisons or by magnitude estimation — the apparent weight of objects having mass m and volume v, and let us suppose that both affect the judgment of weight. If the ordering by apparent weight satisfies the axioms of conjoint measurement and if for some integers p and q Eq. 1 holds — to my knowledge, neither supposition has been tested — then the resulting subjective measure of weight must be the product of two components, each of which is a power function of the extensive physical measure. The fact that S. S. Stevens (1961; and see references given there) and others have repeatedly found (approximate) power relations between (subjective) magnitude estimates and natural physical measures tempts one to investigate these theoretical possibilities further.

Note that if a suitable set of axioms for conjoint measurement are sustained, but Eq. 1 is rejected, then different relations between the two methods of measurement must be considered; of course, these will lead to relations between the conjoint and extensive measures different from Eq. 2.

6. Analysis of Variance and Conjoint Measurement

The additive representation derived in the theory of conjoint measurement is much like the model *postulated* in simple analysis of variance (AOV). It differs from the AOV model in not having either a random error or an interaction term, but in practice these differences are probably not very important. A common null hypothesis of AOV is that the interaction term is zero, and so under that hypothesis the only difference is the random error term. However, as was noted earlier, existing data strongly suggest that purely algebraic measurement theories are inadequate and will have to be replaced by probabilistic generalizations. One of the simplest generalizations is to add a random error term, and such a generalization of conjoint measurement seems to be but a renaming of AOV.

This is not so, and it is important to realize why. In conjoint measurement we construct an additive representation if one exists. In AOV we accept as given some more-or-less arbitrary measure of the attribute of interest and we ask whether, within the variability of the data, this measure is additive over the experimentally independent variables, i.e., we test the null hypothesis that for the given measure there is no interaction. We do not usually attempt to transform the given measure to find the one that is most nearly additive. To the extent that transformations are used in AOV, it is to try to satisfy the equi-normality (normal distributions and homogeneity of variance) assumptions of the statistical test, not to find the "best" additive representation. There is absolutely no reason to expect that the same transformation will fulfill both conditions. Thus, in my opinion, AOV simply does not deal with the often interesting problem of whether the interaction is removeable or inherent : when we conclude from an AOV that an interaction is "significant," we do not draw an absolute conclusion, but only one relative to the given measure.

A trivial example illustrates the point. Suppose that, except for random error, $w = x + y + 2\sqrt{xy}$ and that the random error satisfies the equi-normality assumption. If the data are not too variable, we will conclude from an AOV that $2\sqrt{xy}$ is a significant interaction term. Nevertheless, at the expense of destroying the equi-normality property of the random error, the interaction is completely removed by the square-root transformation since $\sqrt{w} = \sqrt{x} + \sqrt{y}$.

Those who have recognized the relative character of AOV conclusions are cautious to report an additive interaction only when their data completely preclude the possibility of an additive representation. The most commonly accepted evidence for an inherent interaction is data that are non-monotonic, i.e., that "cross" in the sense that, for example, there exist values a and b of the first component and x and y of the second such that both (a, x) P (b, x) and (b, y) P (a, y) are observed. There are more general qualitative conditions that also exclude the possibility of an additive representation. One is a violation of the basic cancellation axiom of conjoint measurement, namely, if (a, x) P (f, s) and (f, r) P (b, x), then (a, r) P (b, s), or, more generally, a violation of any member of Scott's (1964) family of cancellation axioms. Of course, the difficulty with this observation is that random errors make it uncertain whether an observed violation is real or specious. We appear to need a systematic procedure, perhaps a computer program, to find that monotonic transformation of the original measure that, in some appropriate sense, "most nearly" approximates additivity, after which some statistical test should be used to evaluate the null hypothesis that there is no interaction (see Sec. 7).

Before turning to the question of suitable tests, let me remark that some psychologists may attach too much scientific significance to additive independence and so to the existing AOV models. It is no less significant to discover that a measure w can be expressed in terms of measures x and y as $w = xy$, $w = 1/(x + y - xy)^2$, etc., rather than as $w = x + y$. The dependence of w upon the independent variables

x and y is neither less interesting nor more interactive in the first two examples than in the last. The additive model may be a useful starting point when we know very little, but its blind use, leading to the repeated conclusion of statistically significant (additive) interactions, could very well block us from gaining an understanding of which variables control a complex measure and of the mathematical formula for that control. Fortunately, work has now begun on more general measurement models (Tversky, 1967).

7. Hypothesis Testing and Scales of Measurement

Two views have been voiced in the psychological literature about the limitations that the uniqueness of our scales impose upon the statistical inferences we make. One is that the classical parametric tests (which often involve an assumption of normal distributions) are appropriate only for strong — either interval or ratio — scales because relations between certain descriptive statistics (means, variances, etc.) remain invariant only under the admissible transformations of these scales; and for weaker — ordinal — scales we should use weaker — non-parametric or, more precisely, distribution-free — tests. The other view is that these strictures need not be taken seriously because of one or another combination of the following arguments : (a) tests concern distributions of numbers and so it does not matter what type of scale the numbers come from; (b) we usually bring unstated intuitions to bear when choosing numerical assignments for ordinal scales which actually make them much more like interval or ratio scales than we can justify formally; and (c) parametric tests are really quite intensitive to considerable violations of the equi-normality assumptions. The relevant publications are : Anderson (1961), Behan & Behan (1954), Burke (1953), Gaito (1959), Lord (1953), Senders (1953, 1958), Siegel (1956), and Stevens (1946, 1951, 1959).

A third view, somewhat different from either of these, is that the use of a test is limited by the class of transformations under which the null hypothesis is unchanged. This class of transformations will have something to do with the admissible scale transformations if and only if the null hypothesis says something about the scale. Two examples will illustrate what I mean.

First, consider the familiar null hypothesis that two samples of observations come from the same distribution. This hypothesis says nothing at all about the scale used, and as is easily seen it is unchanged by any one-to-one (nominal) scale transformation, because all the transformation does is change the mathematical form of the single distribution that is alleged to account for the data. This invariance holds whether the scale is a ratio scale, such as weight, or an ordinal one, such as I Q. Since any transformation is acceptable, one that makes the distribution normal, if such exists, permits us to use the standard tests; or, equivalently, the inverse transform generates an equivalent test for the original distribution. What is not clear, and to my knowledge has not been investigated, is the effect of such transformations on the power of the test.

Second, consider a null hypothesis that is directly concerned with the scale. Suppose we have provided a criterion for a "best" approximation to additivity in the sense of conjoint measurement and that, for this class of scales, we wish to test the null hypothesis that, within the variability of the data, there is no interaction term. Since we are concerned with no interaction among the "most nearly" additive scales, we are restricted to those transformations that keep us within that class of scales — in this case, linear transformations. It is of no interest at all to show that for some transformation outside this class there is an interaction, which of course there will be even if the null hypothesis is true. Since we cannot expect the equi-normality assumptions of AOV generally to be met and since the nature of the hypothesis precludes a transformation of the scale so that they are met, we are forced either to use a non-parametric test or to show that our data are within the region of insensitivity of a parametric test.

In summary, then, I am suggesting that if the hypothesis under test is not about a property of the scale, we are free to make any transformation we please, in particular, the one that permits us to usė a parametric test. If, however, the hypothesis is about a property of the scale, such as additivity, then we are restricted to the admissible scale transformations and, in general, we will use non-parametric tests for strong scales and parametric ones for weak scales. If so, why then has nearly the opposite been suggested : that strong scales justify the use of strong statistical techniques and weak ones require weaker statistical techniques ? The main reason, I believe, is the fact that the truth of some statements about particular statistics — e.g., that the mean of one group of observations is less than the mean of another — is preserved only when the class of admissible scale transformations is severely limited — in the case of comparisons of means, to linear transformations (interval scales) [1]. In a rough sense, the weaker the scale type,

[1] For a detailed discussion of the invariance and meaningfulness of statements about particular statistics under admissible scale transformations, see Adams, Fagot & Robinson (1965), Stevens (1946, 1951), and Suppes & Zinnes (1963).

the weaker must be a statement about the relation between statistics in order for it to remain true under admissible scale transformations. "In general, the more unrestricted the permissible transformations, the more restricted the statistics." (Stevens, 1959, p. 27). By analogy, I believe, some have concluded that the weaker the scale type, the weaker must be the assumptions of the statistical test used. This, I submit, is the wrong analogy; it is the invariance of the truth of the null hypothesis that limits the acceptable transformations, and this limitation in turn controls our ability to fulfill the assumptions of parametric statistical tests. If the null hypothesis concerns the scale, then and only then do the admissible scale transformation affect the test we choose.

REFERENCES

Aczél, J., Pickert, G., Radó, F. — Nomogramme, Gewebe und Quasigruppen. *Mathematica,* 1960, **2**, 5-24.

Adams, E. W., Fagot, R. F. — A model of riskless choice. *Behav. Sci.,* 1959, **4**, 1-10.

Adams, E. W., Fagot, R. F., Robinson, R. E. — A theory of appropriate statistics. *Psychometrika,* 1965, **30**, 99-127.

Anderson, N. H. — Scales and statistics : parametric and non-parametric. *Psychol. Bull.,* 1961, **58**, 305-316.

Behan, F. L., Behan, R. A. — Football numbers (continued). *Amer. Psychologist,* 1954, **9**, 262-263.

Burke, C. J. — Additive scales and statistics. *Psychol. Rev.,* 1953, **60**, 73-75.

Campbell, N. R. — *Physics : the elements.* Cambridge, Cambridge Univ. Press, 1920. Reprinted as *Foundations of science : the philosophy of theory and experiment.* New York, Dover, 1957.

Campbell, N. R. — *An account of the principles of measurement and calculation.* London, Longmans, Green, 1928.

Cohen, M. R., Nagel, E. — *An introduction to logic and scientific method.* New York, Harcourt, Brace, 1934.

Debreu, G. — Cardinal utility for even-chance mixtures of pairs of sure prospects. *Rev. Econ. Studies,* 1959, **26**, 174-177.

Debreu, G. — Topological methods in cardinal utility theory. In Arrow, K. J., Karlin, S., Suppes, P. (Eds.), *Mathematical methods in the social sciences, 1959.* Stanford, Stanford Univ. Press, 1960, pp. 16-26.

Gaito, J. — Nonparametric methods in psychological research. *Psychol. Rep.,* 1959, **5**, 115-125.

Krantz, D. H. — Conjoint measurement : the Luce-Tukey axiomatization and some extensions. *J. math. Psychol.,* 1964, **1**, 248-277.

LORD, F. M. — On the statistical treatment of football numbers. *Amer. Psychologist,* 1953, **8**, 750-751.

LUCE, R. D. — A "fundamental" axiomatization of multiplicative power relations among three variables. *Philos. Sci.,* 1965, **32**, 301-309.

LUCE, R. D. — Two extensions of conjoint measurement. *J. math. Psychol.,* 1966, **3**, 348-370.

LUCE, R. D., TUKEY J. — Simultaneous conjoint measurement : a new type of fundamental measurement. *J. math. Psychol.,* 1964, **1**, 1-27.

LUCE, R. D., SUPPES, P. — Preference, utility, and subjective probability. In Luce, R. D., Bush, R. R., Galanter, E. (Eds.), *Handbook of mathematical psychology,* Vol. III. New York, Wiley, 1965, pp. 249-410.

MARLEY, A. A. J. — *Some probabilistic models of simple choice and ranking.* Ph. D. dissertation. Philadelphia, University of Pennsylvania, 1965.

McLAUGHLIN, D. H., LUCE, R. D. — Stochastic transitivity and cancellation of preferences between bitter-sweet solutions. *Psychonomic Sc.,* 1965, **2**, 89-90.

ROSKIES, R. — A measurement axiomatization for an essentially multiplicative representation of two factors. *J. math. Psychol.,* 1965, **2**, 266-276.

SCOTT, D. — Measurement structures and linear inequalities. *J. math. Psychol.,* 1964, **1**, 233-247.

SENDERS, VIRGINIA L. — A comment on Burke's additive scales and statistics. *Psychol. Rev.,* 1953, **60**, 423-424.

SENDERS, VIRGINIA L. — *Measurement and Statistics.* New York, Oxford, 1958.

SIEGEL, S. — *Non-parametric statistics.* New York, McGraw-Hill, 1956.

STEVENS, S. S. — On the theory of scales of measurement. *Science,* 1946, **103**, 667-680.

STEVENS, S. S. — Mathematics, measurement, and psychophysics. In Stevens S. S. (Ed.) *Handbook of experimental psychology.* New York, Wiley, 1951, pp. 1-49.

STEVENS, S. S. — Measurement, psychophysics, and utility. In Churchman, C. W., Ratoosh, P. (Eds.). *Measurement : definitions and theories,* New York, Wiley, 1959, pp. 18-63.

STEVENS, S. S. — The psychophysics of sensory function. In Rosenblith, W. A. (Ed.), *Sensory Communication.* New York, Wiley, 1961, pp. 1-33.

SUPPES, P. — A set of independent axioms for extensive quantities. *Portugaliae Mathematica,* 1951, **10**, 163-172.

SUPPES, P., ZINNES, J. L. — Basic measurement theory. In Luce, R. D., Bush, R. R., Galanter, E. (Eds.), *Handbook of mathematical psychology,* Vol. I. New York, Wiley, 1965, pp. 1-76.

TVERSKY, A. — *Finite additive structures.* Michigan Mathematical Psychology Program Technical Report MMPP 64-6. Ann Arbor, University of Michigan, 1964.

TVERSKY, A. — *Additive analyses choice behavior : a test of utility theory.* Michigan Mathematical Psychology Program Technical Report MMPP 65-2. Ann Arbor, University of Michigan, 1965.

TVERSKY, A. — A general theory of conjoint measurement. *J. math. Psychol.,* 1967, **4**, in press.

DISCUSSION

Rouanet. — Pourriez-vous nous donner quelques détails sur les « hypothèses techniques » que vous avez mentionnées au moins deux fois ?

Luce. — You mean the assumptions underlying conjoint measurement ? In the presentation that Tukey and I gave, (Luce and Tukey, 1964), there are four axioms. Let R be a relation on $A_1 \times A_2$. The first axiom is that R is a weak ordering of $A_1 \times A_2$. The second axiom is the following cancellation property : if (a, x) R (f, s) and (f, r) R (b, x), then (a, r) R (b, s). The third axiom is known as the solution-of-equations property : given a and b in A_1 and x in A_2, there exists some y in A_2 such that (a, x) I (b, y). And the fourth, which I will not try to state exactly, is an Archimedean condition. It requires a definition before one can state it, but it is a fairly standard type of axiom. Now, the thing closest to continuity — and it is not strictly a continuity assumption — is the solution-of-equations axiom. The drawback of that axiom, as stated in our paper, is that it is postulated to hold without restriction, and this means that the scales have to be unbounded. In general, this is not likely to be true for psychological variables. As I indicated earlier, I have recently (Luce, 1966) weakened this axiom considerably at the expense of having to add, mainly, one other axiom, namely, that for all x and y, (a, x) R (b, y) holds if, and only if, (a, y) R (b, y) holds. In the original system, we were able to deduce this property; however, once the unrestricted solution of equations is dropped, it can no longer be deduced and so it has to be added as a separate axiom. In addition, there is another axiom that asserts the existence of a sufficient number of elements, but I don't think it is worth stating it explicitly.

Audley. —It is unclear to me how wide the range of transformations on the null hypothesis could be. In order to carry out parametric statistics one must know that the populations under examination have certain characteristics, and it is not often that one has enough information from the sample to know what kind of transformation should be employed to achieve, say, approximate normality.

Luce. — I certainly agree that the problem of deciding exactly which transformation should be used is quite difficult; I have not addressed myself to that question. Rather I have considered whether or not it is acceptable to make arbitrary transformations when you are testing the null hypothesis that the two samples came from the same distribution. Stevens has argued that there are limitations which are based on the scale type of the measurements, and I am arguing, not so : that if the null hypothesis says that the two samples came from the same distribution, one may take any transformation one wishes. Then if we could solve the problem of which transformation converts the given distribution into a normal one, then I would think that we would want to use it. But I did not say how to find this transformation. All I said is that we are free to make the transformations on the data because, so far as I can see, all we are alleging is that the two samples came from a common distribution, and that will remain true no matter how we transform the measure. On the other hand, if a null hypothesis concerns the nature of the scale, then I think we are limited in the admissible transformations. It seems to me that an analysis of variance differs significantly from testing a null hypothesis such as the one just discussed because it is concerned with the nature of the scale; namely, whether the scale exhibits additivity over the coordinates. I think that these are two quite different classes of null hypotheses, and the effects of the scale transformations are totally different in the two situations. I feel that this has been fairly thoroughly muddled in the literature. But as to which transformation to use in order to justify the use of a parametric test is a problem beyond my competence; it is a statistical question that I am not prepared to deal with here and probably not anywhere.

AUDLEY. — May I pursue the point a little more; I am not sure it is just a statistical matter. In order to be able to make a statistical inference about the population from sample information, it seems to me that any transformations used will probably have to be of a kind that preserve some properties of the scale type.

LUCE. — If you are correct in that supposition, then I would agree that the scale type will make a difference. However, I think that it is doubtful that this is a generally true statement. I can imagine that if you had an adequate theory for the situation with which you are dealing, then the theory might very well tell you what transformations you should use and so I can see that the situation you describe might arise. But I can also imagine situations where the scale type simply would not matter. You would like to use your data to infer what transformation produces approximate normality. A series of approximations would be involved here, which, no doubt, will complicate considerably the statistical inference problem.

KLIX. — It seems to me that we have here a very meaningful way for getting an applied form of measure theory, adapted to problems relevant in psychological research. But in connection with your explanation I become aware of several difficulties. One has been mentioned already. I mean the difficulties which arise with statistical features of the outcomes in psychological measurement. I would like to mention one difficulty in this connection : it is quite certain that the two parameters do not have the same distribution. On the contrary, examples could be given where the probability distribution over the conjoint parameters are quite different.

And my second problem : do you intend to unfold the two-component condition to three or more components ? There are many examples in psychological research of inferences caused by more than two conditions. For instance, in perceptual theory the apparent velocity (as a comparable and therewith measurable unit) is caused by more than three conditions : (1) the real or physical speed, (2) the intensity of the stimulus, (3) its physical distance, (4) the neighbourhood, and others. It seems to me that difficulties arise if your conjoint approach is applied to more than two conditions. There exists no symmetry in the weights of the components, their effects on the inference are not homogeneous and they are interchangeable in a limited, condition-dependent degree. Do you believe these difficulties are to be mastered ?

LUCE. — I certainly agree that the statistical problem seems, right now, to be the most important next step to be pursued. There are, indeed, situations for which the distributions on the two coordinates are quite different. For example, loudness depends both on intensity and frequency, but the dependence is very, very different on the two coordinates, and the statistical features are therefore very different. About the second point — the possibility of more than two coordinates — it is fairly straight-forward to generalize the additive theory to any finite number of coordinates. Krantz (1964) has given one such generalization, and I have given a somewhat weaker generalization (Luce, 1966), so on this point we do not disagree. There is really no serious problem in generalizing additivity to *n*-coordinates. However, when additivity is abandoned — when some other functional relationship is assumed — matters get a bit more complicated. Tversky (1967) has given very abstract, necessary and sufficient conditions for polynomial representations on a finite number of coordinates. The translation of these conditions into testable hypotheses appears to be quite difficult. The conditions are stated quite abstractly and really involve an infinite set of conditions although they are written so that there appear to be only two axioms. But when you disentangle them it turns out that one is really an infinity of axioms. From the point of view of the experimentalist, it is not very clear what to do with axioms of this sort. The reason that Tukey and I and others have been concerned with sufficient conditions, rather than necessary and sufficient conditions, has been to find systems that are potentially testable in the laboratory. Basically, we have shown that only one or two cancellation properties are needed provided that some sort of solution-of-equations condition is satisfied.

There is no question — and let me not be misinterpreted on this point — that there are any number of situations to which this type of theory is not applicable. The hope is that there are a few to which it is applicable. But at the moment I do not believe that anybody knows whether such situations actually exist in psychology. I think that we have to take the point of view that this additive theory is, perhaps, the simplest of a set of possible measurement theories — the simplest one that might conceivably work — and now the problem is to go into the laboratory, to try various likely empirical candidates, and to see if, in fact, it works anywhere. It may very well not work anywhere. At the moment no one is doing much experimental research on the problem because the statistical problems have not been resolved. I know of no experiment proposed to test these theories in which we would not anticipate inconsistent data from the subjects. The attempts that I have made, both on loudness and on taste, have produced probabilistic data. Of course, it may just be that we don't know how to do the experiments properly or it may be inherent in the subjects. To be sure, physicists have had the same difficulty, and they have tended to handle it in a fairly casual manner. They simply assume that there is a little random error scattered around in such a way that by making enough observations and taking means everything is alright. They seem to get along pretty well doing this; whether we can manage the same thing is not so clear.

Suppes. — It seems to me that there are some genuine differences between the interest in measurement in psychology and in physics. The physicists working in many domains, at least experimentally, work under the hypothesis that it is natural to increase the accuracy of the measurements. Unless we go on to far more mechanical or atomic domains, where this is a general methodological postulate, it is not a postulate of interest for much psychological research. When you are working with the human observer you aren't really concerned necessarily, it seems to me, in many investigations involving the measurement of human skills, attitudes, judgements, etc. to have a theory that leads to the increasing accuracy of the measurements; and this makes the considerations rather different in psychology than in physics.

Luce. — It is not too often that Dr. Suppes and I differ on questions of measurement — we have talked a lot about them — but I guess I disagree with him to some extent here. It does not seem *a priori* clear to me that we cannot increase the precision of our measurements in psychology. We take for granted certain experimental procedures which are, by now, classical, and we live with them, but these procedures may produce certain inaccuracies. Whether we can rid ourselves of these inaccuracies can not really be prejudged. Let me cite an example in a nearby area — not actually in measurement, as such. A graduate student and I have been doing some work on reaction time, and we keep modifying our physical measurements of reaction time in an attempt to find out how much of the total variability is contributed by the subject and how much by our apparatus. The latter contribution is not inconsiderable, and it is fairly tricky to eliminate it from the apparatus. We are using what is considered relatively good equipment; nevertheless, it looks as though something of the order of 50 % of the variability we observe is due to the equipment and not to the subject. Don't hold me to that figure; we are still in the process of trying to find out just how much is due to the apparatus. When, as we can easily achieve with simple reaction times, 60 % of the observations fall within 20 ms. band, then a 5 ms. variation somewhere in the apparatus constitutes a significant part of the variability. One cannot but wonder about the degree to which it will be possible in other areas to reduce variability by increased care of experimentation. In the area of measurement, I wonder if we were to modify our procedures appreciably, could we get rid of more and more of the variability. I would hate to prejudge the answer. Some of the probability models we build may not be models of the behavior of human beings but of the inadequacies of our experimental procedure and apparatus. This possibility frequently haunts me as we construct ever more probability models.

Suppes. — It's fun to pursue this for a moment. I accept certainly what you say about reaction times, but it seems to me that for many areas of behavior one can certainly make a case, let's say, that the behavior of the human being is rough and ready and it is not at all clear that we want, or are interested in, a theory of measurement of that behavior that has the same kind of refinements as in physics. For example, the kind of skills involved in walking up stairs, hitting a tennis ball, making a decision in terms of maximizing some property. I mean this is at least one way of looking at much behavior. There is a kind of robustness and at the same time a kind of crudity in the measurements used by the organism in taking a decision or making a response.

Luce. — I agree with that.

Rouanet. — A propos du test de l'additivité, vous recommandez, si je comprends bien, pour tester l'interaction, d'utiliser un test non-paramétrique une fois que l'échelle a été ajustée le mieux possible. Avez-vous des recommandations plus précises et suggérez-vous des tests non-paramétriques particuliers ?

Luce. — No. Not being a statistician I have some difficulty in answering this, but my impression is that we do not have, at the moment, suitable test procedures to analyze this problem. What I am indicating, essentially, is a problem that I think needs to be solved, but I am not proposing any explicit solution. What concerns me is that statisticians seem to act as if, and many psychologists seem to agree with them that, the analysis of variance, as we now know it, solves the problem. And I am saying, no, the real problem of finding additive measures must be dealt with somewhat differently, but I do not have any explicit proposals as to how it should be done. There are two aspects to this problem and I do not have explicit proposals for either part. First, how does one find the "best" additive representation ? Second, given the solution to that, how does one do the statistical test without violating the additivity requirement ?

ROUANET. — Je songeais aux tests non-paramétriques, du genre « tests de permutation » de Fisher, qui, dans bien des cas, donnent des résultats voisins de ceux obtenus par le F classique de l'analyse de variance, tout en se passant de l'hypothèse de l'équinormalité. D'autre part, ces tests ont le mérite d'être relatifs à une échelle d'intervalles, et non à une échelle de type inférieur, ainsi ils sont aussi puissants que les tests paramétriques habituels.

LUCE. — I think this is the familiar question : can you get away with a parametric test when the assumptions of that test are not completely fulfilled ? You know as much about this as I do. There is a certain amount of semi-empirical literature about it to the effect that one can get away with it provided that the assumptions are not too badly violated. Of course, all these considerations will apply here : we will, in many situations, be able to use a parametric test without serious error. Nevertheless, in some situations this may not be true, and one would like to have a non-parametric test available. I think, though, the important point is not this old question about when you can use parametric tests, but rather the need for transformations to get the best additive measure. Psychologists continually use the analysis of variance to conclude the existence of interaction, and we — in particular, a group such as this — have got to fight the common misinterpretations. I don't think most psychologists really know what they are doing when they test for interaction in analysis of variance; they are not concluding a significant interaction, as they believe, but rather a significant interaction relative to the particular, often arbitrary, measure that they have used. There may be no interaction at all in a different monotonic measure.

SIMON. — I wonder whether you don't need to add a third class of difficulties to the difficulties of statistical testing that have been mentioned here. The leading statistical theorists, I believe, are quite dissatisfied with the theory of testing extreme hypotheses. By an extreme hypothesis I mean an hypothesis in which the null hypothesis is a very specific theory or model, and the alternative is its denial. Now the hypotheses you are testing in your scaling work, as well as most of the models that are described in other papers in this conference, are extreme hypotheses in this sense.

The inapplicability of standard hypothesis testing methods to extreme hypotheses is well known. Any of the classical physical theories which are well established as very good first approximations to the world — the law of falling bodies, for example — will be rejected by tests of extreme hypotheses if the data are good enough and the samples large enough. On the other hand, when you use non-parametric tests here, you are guaranteeing the acceptance of your hypotheses by using tests of extremely low power.

Luce. — Certainly there is a danger of this, and I think all the comments you have made are quite relevant. We are just going to have to be very careful not to accept a null hypothesis artificially through the use of weak tests. I do not want to get into a position of arguing against what I think is correct, such as the point just made by Dr. Simon. The point that I have been trying to get at had to do with the interlock between scales of measurement and the use of tests, about which there has been a moderate amount of discussion in the psychological literature and a lot of heated debate. The position I am taking is that neither side of the debate as it now exists in the literature is correct, that there is actually a third position which is that you are not limited in applying transformations provided that you are not making a test about the nature of the scale. But if you are making such a test about the scale, then you may be limited in principle subject, however, to all the caveats about the use of parametric tests when the assumptions are not strictly satisfied and about being careful not to accept the null hypothesis because too few data were collected for the strength of the test. I agree, but the point I was trying to make was a little different than that. I did not want to get into the whole problem of doing statistical tests; I am certainly not the person to discuss that.

Audley. — Might we explore a little the relation between measurement and theory in general ? In the examples of measurement you have discussed, we are dealing with a fairly closed system in which the theoretical structure is well known. The measurement scheme is then worked out within this structure. In many psychological situations the selection of variables for consideration may be more arbitrary. In your example of apparent weight, just two particular variables were selected. But one feels that these are only a sample of all the variables involved in the phenomenon. What can be done about building up systems of measurement when we do not know all the major variables entering into a situation ?

Luce. — My general impression is that there is very little a formal theory can do to help here. The discovery of relevant variables and getting data about their interaction with one another does not seem to be something you can reduce to a routine. This is the classic problem of the factor analysis : can't we be rid of the difficult and grubby problem of finding out how the world is put together and do it all formally ? Well, from my point of view, factor analysis has pretty much run its course in psychology, although, unfortunately, I see signs of it moving into the physical sciences. Somewhere I recently read a geologist who recommended the use of factor analysis in geology; I can only feel sorry for his discipline if he is heeded. I just don't think that this kind of pure formalism provides any hope of unlocking the variables. To do that requires whatever it is that we, as scientists, provide : insight and creativity of various sorts. All that a measurement theory of the type I have been talking about can do is to provide a possible framework in which the data might fit; and if one finds that some variables do in fact fit it, then we will begin to be in a position to see which variables are fundamental. If nothing fits it, then it is one of those trys that didn't work. But I don't think there is any nice, sweet avenue to the discovery of the relevant variables. Certainly, no purely formal procedures are going to help.

Restle. — I don't quite understand the relation between the two variables a and x that you describe. I take it they must be different, yet both be attributes of the same object, the object being measured. Furthermore, we are abstracting from these two attributes some common property. I have done some work on judgment of complex objects, and it is known that one cannot compare the brightness of a blue and yellow light as accurately as two blue lights. To apply conjoint measurement you must have complex objects, and I wonder what is the psychological significance of the complexity of objects in conjoint measurement ?

Luce. — I think we must be careful in the choice of stimuli to which we attempt to apply conjoint measurement. I know the problem you are driving at. It appears, however, that for certain judgmental variables more than one physical variable affects the psychological measure. An example is the apparent weight of objects, which, on the one hand, seems to be a reasonably unitary psychological notion and, on the other, can be manipulated to some degree both by the mass of the object and by its volume or, equivalently, its density. Another example is the loudness of pure tones which depends both on intensity and frequency, albeit only slightly on frequency. It is in situations of this sort that one might hope that the axioms of some conjoint measurement theory would be satisfied and so the resulting representation could be used. If you attempt to apply these theories to truly complex stimuli, my guess is that you are asking for trouble at this stage of the game. Rather, I am inclined to try to apply them to what appear to be psychologically unitary variables that happen to depend on more than one physical attribute.

Restle. — The contrasting approach would be to take up your weight-lifting example as the study of an illusion, and your problem on loudness of tones as one of comparing the loudness of two tones having different pitch. I suggest that the conjoint measurement theory may not take sufficient cognizance of the perceptual situations, hence may be irrelevant to understanding such experiments. I do not question the formal importance and clarity of the development as you do it.

Luce. — You may be right. I do not know how to decide such questions in advance, and it may turn out in the end that the only interest in this kind of formal development is philosophical. I think that it is of interest for, at least, the philosophy of physics, but it may not go beyond that. I am perfectly prepared to see that happen, although I rather hope that it doesn't.