Meaningful and Meaningless Statements in Epidemiology and Public Health

by

Fred S. Roberts

DIMACS Center

Rutgers University

Piscataway, NJ 08854 USA

Acknowledgements: The author gratefully acknowledges the support of the National Science Foundation under grant number EIA-02-05116 to Rutgers University. A number of ideas and some of the examples and language in this paper are borrowed from my paper Roberts (1994), which explores meaningful and meaningless statements in operations research.

1 Introduction

The theory of measurement is an interdisciplinary subject that grew out of the attempt to put the foundations of measurement on a firm mathematical foundation. Building on classic examples of measurement in the physical sciences, the theory was motivated by the attempt to make measurement in economics, psychology, and other disciplines more precise. The theory traces its roots to work of Helmholtz (1887/1930), and was widely formalized in the 20th century in such books as Krantz, Luce, Suppes, and Tverksy (1971), Luce, Krantz, Suppes, and Tversky (1990), Pfanzagl (1968), Roberts (1979/2009), and Suppes, Krantz, Luce, and Tversky (1989). Measurement theory is now beginning to be applied in a wide variety of new areas. Little known in the fields of epidemiology and public health, the theory has the potential to make important contributions to epidemiological measurement. In turn, problems of epidemiology are posing new challenges for measurement theory.

We will seek to answer questions such as the following:

- Is it meaningful to say that the malaria parasite load has doubled?
- Is the average cough score for one set of TB patients higher than that for another?
- For controlling the spread of HIV, which of abstinence education, universal screening, and condom distribution are more effective?

All of these questions have something to do with measurement. We will provide a brief introduction to the theory of measurement, with an emphasis on the types of scales that can arise.

In almost every practical application in epidemiology, something is measured. Yet in many cases, little attention is paid to the limitations that the scales of measurement being used might place on the conclusions that can be drawn from them. Scales may to some degree be arbitrary, involving choices about zero points or units or the like. It would be unwise to make decisions that could turn out differently if the arbitrary choice of zero point or unit is changed in some "admissible" way. We shall be interested in exploring the extent to which this can happen, and in laying down guidelines for what conclusions based on scales of measurement are allowable. We will make these ideas precise by introducing a key measurement theory concept of meaningfulness. Using examples from the study of diseases such as HIV, malaria, and tuberculosis, we will give a variety of examples of meaningless and meaningful statements. More subtle applications will include averaging judgments of cough severity or judgments of fatigue, finding measures of air pollution combining different pollutants, and evaluating alternative HIV treatments using "merging" procedures for normalized scores. We will also discuss the meaningfulness of statistical tests and of answers to optimization questions in epidemiology arising from problems such as the effect of climate change on health. We will then discuss general results about how to average scores to attain measures allowing meaningful comparisons and close with a discussion about measurement issues arising in the study of behavioral responses to health events.

2 Scales of Measurement

It seems clear that measurement has something to do with numbers. In this paper, it will suffice to think of assigning real numbers to objects. Our approach to scales of measurement is based on the notion, going back to the psychologist S.S. Stevens (1946,1951,1959), that the properties of a scale are captured by studying *admissible transformations*, transformations that lead from one acceptable scale to another. For example, we transform temperature measurements from Centigrade into Fahrenheit and mass measurements from kilograms into pounds. Assuming a scale assigns a real number $f(a)$ to each object a being measured, an admissible transformation of scale can be thought of as a function ϕ that takes $f(a)$ into $(\phi \circ f)(a)$.

Our approach to scales of measurement is based on ideas introduced by Stevens. Assuming that a scale assigns a real number to each object being measured, we call a scale a *ratio scale* if the admissible transformations are of the form $\phi(x) = \alpha x$, $\alpha > 0$, an *interval scale* if the admissible transformations are of the form $\phi(x) = \alpha x + \beta$, $\alpha > 0$, an *ordinal scale* if the admissible transformations are the (strictly) monotone increasing transformations, and an *absolute scale* if the only admissible transformation is the identity. For definitions of other scale types, see Roberts (1979/2009). Thus, in the case of ratio scales, the scale value is determined up to choice of a unit; in the case of interval scales, it is determined up to choices of unit and of zero point; and in the case of ordinal scales, it is determined only up to order. Mass is an example of a ratio scale. The transformation from kilograms into pounds, for example, involves the admissible transformation $\phi(x) = 2.2x$. Length (inches, centimeters) and time intervals (years, seconds) are two other examples of ratio scales. It is sometimes argued that scales developed for loudness ("sones") define a ratio scale, but this is not universally accepted. Temperature (except where there is an absolute zero) defines an interval scale. Thus, transformation from Centigrade into Fahrenheit involves the admissible transformation $\phi(x) = (9/5)x + 32$. Time on the calendar is another example of an interval scale (to say that this is the year 2011 involves not only a choice of unit but a choice of the zero year). An example of an ordinal scale is any scale in which we give grades of materials, such as leather, lumber, wool, etc. Expressed preferences sometimes lead only to ordinal scales, if we know our preferences only up to order. Subjective judgments in epidemiology, such as of cough or fatigue, often define ordinal scales. Counting gives an example of an absolute scale.

For many scales, the scale type can be determined by showing that the scale arises from a (numerical) representation. In the rest of this paragraph, we say a word or two about how this is done in the theory in such books as Krantz, et al. (1971), Pfanzagl (1968), and Roberts (1979/2009), though the details will not be needed for what follows. Specifically, one studies certain *observed relations* on a set of objects of interest, relations such as "a is longer than b," "a is louder than b," "I think the value of b is between the value of a and the value of c," etc. One identifies corresponding *numerical relations*, relations on a set of real numbers, for instance the "greater than" relation or the "betweenness" relation. Then one studies mappings that take each object of interest into a number so that objects related in a certain way in an observed relation correspond to numbers related in the same way in the corresponding numerical relation. For example, one seeks to assign numbers to objects so that a is judged louder than b if and only if the number assigned to a is greater than the number assigned to b. Such a mapping from objects to numbers is called a *homomorphism* from the observed relation to the numerical relation. In measurement theory, scales are identified with homomorphisms. Formally, an *admissible transformation* of a scale is then a transformation of the numbers assigned so that one gets another homomorphism. In some cases, one can derive a characterization of the class of admissible transformations by working from a (numerical) representation. For details on how to formalize these ideas, see Roberts (1979/2009).

It should be remarked that many scales based on subjective judgments cannot be derived from a (numerical) representation. Then, we must use the principle that the admissible transformations are those that preserve the information carried by the scale. Knapp(1990) and Thomas (1985) emphasize the difficulties involved in identifying scale type. As Stevens (1968) argues, it is often a matter of empirical judgment to determine the admissible transformations and hence the scale type.

3 Meaningful Statements

In measurement theory, we speak of a statement as being meaningful if its truth or falsity is not an artifact of the particular scale values used. The following definition is due to Suppes (1959) and Suppes and Zinnes (1963).

Definition: *A statement involving numerical scales is meaningful if its truth or falsity is unchanged after any (or all) of the scales is transformed (independently?) by an admissible transformation.*

A slightly more informal definition is the following:

Alternate Definition: *A statement involving numerical scales is meaningful if its truth or falsity is unchanged after any (or all) of the scales is (independently?) replaced by another acceptable scale.*

In some practical examples, for example those involving preference judgments or judgments of "louder than" under the "semiorder" model (Roberts, 1979/2009, 1994), it is possible to have scales where one cannot go from one to the other by an admissible transformation, so one has to use this alternate definition.

We will avoid the long literature of more sophisticated approaches to meaningfulness. Situations where this relatively simple-minded definition may run into trouble will be disregarded. Emphasis is to be on applications of the invariance motivation behind the theory of meaningfulness.

Consider the following statement:

Statement S: "The duration of symptoms in an influenza victim not treated with Tamiflu is three times as long as the duration of symptoms in an influenza victim who is so treated."

Is this meaningful? We have a ratio scale (time intervals) and we consider the statement:.

$$f(a) = 3f(b). \tag{1}$$

This is meaningful if f is a ratio scale. For, an admissible transformation is $\phi(x) = \alpha x, \alpha > 0$. We want Equation (1) to hold iff

$$(\phi \circ f)(a) = 3(\phi \circ f)(b). \tag{2}$$

But Equation (2) becomes

$$\alpha f(a) = 3\alpha f(b) \tag{3}$$

and (1) iff (3) since $\alpha > 0$. Thus, the statment S is meaningful.

Consider next the statement:

Statement T "The patient's temperature at 9AM today is 2 per cent higher than it was at 9 AM yesterday."

Is this meaningful? This is the statement

$$f(a) = 1.02f(b).$$

This is meaningless. It could be true with Fahrenheit and false with Centigrade, or vice versa.

In general, for ratio scales, it is meaningful to compare ratios:

$$f(a)/f(b) > f(c)/f(d).$$

For interval scales, it is meaningful to compare intervals:

$$f(a) - f(b) > f(c) - f(d).$$

For ordinal scales, it is meaningful to compare size:

$$f(a) > f(b).$$

Let us consider another example. Malaria parasite density is mainly obtained by reading slides under microscopes. Consider the statement:

Statement M "The parasite density in this slide is double the parasite density in that slide."

Is this meaningful? Density is measured in number per microliter. So, if one slide has 100,000 per μL and another 50,000 per μL, is it meaningful to conclude that the first slide has twice the density of the second? This is meaningful. Volume involves ratio scales and counts are absolute scales. However: This disregards errors in measurement. A statement can be meaningful in the measurement theory sense but meaningless in a practical sense.

Here is still another example:

Statement W "The second tumor weighs 20 million times as much as the first one."

This is meaningful. It involves ratio scales. It is surely false no matter what the unit. Note that meaningfulness is different from truth. It has to do with what kinds of assertions it makes sense to make, which assertions are not accidents of the particular choice of scale (units, zero points) in use.

4 Averaging Judgments of Cough Severity

Suppose we study two groups of patients with tuberculosis. Let $f(a)$ be the cough severity of a as judged on one of the subjective cough severity scales in use (e.g., rate severity as 1 to 5). Suppose that data suggests that the average cough severity for patients in the first group is higher than the average cough severity of patients in the second group. Is this meaningful?

Let $a_1, a_2, \ldots a_n$ be patients in the first group and $b_1, b_2, \ldots, b_m$ be patients in the second group. Note that m could be different from n. Then we are (probably) asserting that

$$\frac{1}{n}\sum_{i=1}^{n} f(a_i) > \frac{1}{m}\sum_{i=1}^{m} f(b_i). \tag{4}$$

We are comparing arithmetic means. The statement (4) is meaningful if and only if under admissible transformation ϕ, (4) holds if and only if

$$\frac{1}{n}\sum_{i=1}^{n}(\phi \circ f)(a_i) > \frac{1}{m}\sum_{i=1}^{m}(\phi \circ f)(b_i) \tag{5}$$

holds. If cough severity defines a ratio scale, then (5) is the same as

$$\frac{1}{n}\sum_{i=1}^{n}\alpha f(a_i) > \frac{1}{m}\sum_{i=1}^{m}\alpha f(b_i), \tag{6}$$

for some positive α. Certainly (4) holds if and only if (6) does, so (4) is meaningful.

Note that (4) is still meaningful if f is an interval scale. For instance, we could be comparing temperatures. It is meaningful to assert that the average temperature of the first group is higher than the average temperature of the second group. To see why, note that (4) is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n}[\alpha f(a_i) + \beta] > \frac{1}{m}\sum_{i=1}^{m}[\alpha f(b_i) + \beta],$$

where $\alpha > 0$.

However, (4) is easily seen to be meaningless if f is just an ordinal scale. To show that comparison of arithmetic means can be meaningless for ordinal scales, note that we are asking experts for a subjective judgment of cough severity. It seems that $f(a)$ is measured on an ordinal scale, e.g., 5-point scale: 5=extremely severe, 4=very severe, 3=severe, 2=slightly severe, 1=no cough. In such a scale, the numbers may not mean anything; only their order matters. Suppose that group 1 has three members with scores of 5, 3, and 1, for an average of 3, while group 2 has three members with scores of 4, 4, and 2 for an average of 3.33. Then the average score in group 2 is higher than the average score in group 1. On the other hand, suppose we consider the admissible transformation ϕ defined by $\phi(5) = 100$, $\phi(4) = 75$, $\phi(3) = 65$, $\phi(2) = 40$, $\phi(1) = 30$. Then after transformation, members of group 1 have scores of 100, 65, 30, with an average of 65, while those in group 2 have scores of 75, 75, 40, with an average of 63.33. Now, group 1 has a higher average score. Which group had a higher average score? The answer clearly depends on which version of the scale is used. Of course, one can argue against this kind of example. As Suppes (1979) remarks in the case of a similar example having to do with grading apples in four ordered categories, "surely there is something quite unnatural about this transformation" ϕ. He suggests that "there is a strong natural tendency to treat the ordered categories as being equally spaced." However, if we require this, then the scale is not an ordinal scale according to our definition. Not every strictly monotone increasing transformation is admissible. Moreover, there is no reason, given the nature of the categories, to feel that this is demanded in our example. In any case, the argument is not with the precept that we have stated, but with the question of whether the five point scale we have given is indeed an ordinal scale as we have defined it. To complete this example, let us simply remark that comparison of medians rather than arithmetic means is meaningful with ordinal scales: The statement that one group has a higher median than another group is preserved under admissible transformation.

Similar considerations apply to measuring of average fatigue. Fatigue is an important variable in measuring the progress of patients with serious diseases. One scale widely used in measuring fatigue is the Piper Fatigue Scale. It asks questions like: On a scale of 1 to 10, to what degree is the fatigue you are feeling now interfering with your ability to complete your work or school activities (1 = none, 10 = a great deal)? On a scale of 1 to 10, how would you describe the degree of intensity or severity of the fatigue which you are experiencing now (1 = mild, 10 = severe)? A similar analysis applies: It is meaningless to compare arithmetic means, meaningful to compare medians.

Let us return to cough severity, but now suppose that each of n observers is asked to rate each of a collection of patients as to their relative cough severity. Alternatively, suppose we rate patients on different criteria or against different benchmarks. (A similar analysis applies with performance ratings, importance ratings, etc.) Let $f_i(a)$ be the rating of patient a by judge i (or under criterion i). Is it meaningful to assert that the average rating of patient a is higher than the average rating of patient b? A similar question arises in fatigue ratings, ratings of brightness of rash, etc. We are now considering the statement

$$\frac{1}{n}\sum_{i=1}^{n} f_i(a) > \frac{1}{n}\sum_{i=1}^{n} f_i(b). \tag{7}$$

Note in contrast to statement (4) that we have the same number of terms in each sum and that the subscript is now on the scale value f rather than on the alternative a or b. If each f_i is a ratio scale, we then ask whether or not (7) is equivalent to

$$\frac{1}{n}\sum_{i=1}^{n} \alpha f_i(a) > \frac{1}{n}\sum_{i=1}^{n} \alpha f_i(b),$$

$\alpha > 0$. This is clearly the case.

However, we have perhaps gone too quickly. What if f_1, f_2, ..., f_n have independent units? In this case, we want to allow independent admissible transformations of the f_i. Thus, we must consider

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_i f_i(a) > \frac{1}{n}\sum_{i=1}^{n}\alpha_i f_i(b), \tag{8}$$

all $\alpha_i > 0$. It is easy to find $\alpha_i^{\prime}s$ for which (7) holds but (8) fails. Thus, (7) is meaningless. Does it make sense to consider different α_i? It certainly does in some contexts. Consider the case where the alternatives are animals and one expert measures their improved health in terms of their weight gain while a second measures it in terms of their height gain.

The conclusion is that we need to be careful when comparing arithmetic mean ratings, even when we are using ratio scales. Norman Dalkey [personal communication] was the first person to point out to the author that, in many cases, it is safer to use geometric means, a conclusion which by now is "folklore." For consider the comparison

$$\sqrt[n]{\prod_{i=1}^{n} f_i(a)} \quad > \quad \sqrt[n]{\prod_{i=1}^{n} f_i(b)}. \tag{9}$$

If all $\alpha_i > 0$, then (9) holds if and only if

$$\sqrt[n]{\prod_{i=1}^{n} \alpha_i f_i(a)} \quad > \quad \sqrt[n]{\prod_{i=1}^{n} \alpha_i f_i(b)}.$$

Thus, if each f_i is a ratio scale, then even if experts change the units of their rating scales independently, the comparison of geometric means is meaningful even though the comparison of arithmetic means is not. An example of an application of this observation is the use of the geometric mean by Roberts (1972, 1973). The problem arose in a study of air pollution and energy use in commuter transportation. (Health effects of air pollution will be discussed in the next section.) A preliminary step in the model building involved the choice of the most important variables to consider in the model. Each member of a panel of experts estimated the relative importance of variables using a procedure called magnitude estimation. (Here, the most important variable is given a score of 100, a variable judged half as important is given a score of 50, and so on.) There is a strong body of opinion that magnitude estimation leads to a ratio scale, much of it going back to Stevens. (See the discussion in Roberts (1979/2009, pp. 179-180).) How then should we choose the most important variables? By the discussion above, it is "safer" to combine the experts' importance ratings by using geometric means and then to choose the most important variables as those having the highest geometric mean relative importance ratings, than it is to do this by using arithmetic means. That is why Roberts (1972, 1973) used geometric means.

5 Measurement of Air Pollution

There is a close relationship between pollution and health. Various pollutants are present in the air. Some are carbon monoxide (CO), hydrocarbons (HC), nitrogen oxides (NOX), sulfur oxides (SOX), and particulate matter (PM). Also damaging are products of chemical reactions among pollutants. For example, oxidants such as ozone are produced by HC and NOX reacting in the presence of sunlight. Some pollutants are more serious in the presence of others, e.g., SOX are more harmful in the presence of PM. In the early days of air pollution science, there was an attempt to find a way to measure pollution with one overall measure. To compare pollution control policies, we need to compare effects of different pollutants. We might allow increase of some pollutants to achieve decrease of others. One single measure could give indication of how bad the pollution level is and might help us determine if we have made progress. A simple approach is to combine the weight of pollutants. Let us measure the total weight of emissions of pollutant i over a fixed period of time and sum over i. Let $e(i,t,k)$ be the total weight of emissions of pollutant i (per cubic meter) over the tth time period and due to the kth source or measured in the kth location.

Then our pollution index is simply

$$A(t,k) = \sum_{i=1}^{n} e(i,t,k),$$

if there are n pollutants under consideration.

Using this measure, Walther (1972) reached such conclusions as: (i) The largest source of air pollution is transportation and the second largest is stationary fuel combustion (especially by electric power plants); ii) transportation accounts for over 50 per cent of all air pollution; (iii) CO accounts for over 50 per cent of all emitted air pollution. Statement (i) is just

$$A(t,k) > A(t,k').$$

Statement (ii) is just the statement that

$$A(t, k_r) > \sum_{k \neq k_r} A(t, k).$$

Statement (iii) is just the statement that

$$\sum_{t,k} e(i, t, k) > \sum_{t,k} \sum_{j \neq i} e(j, t, k).$$

All these conclusions are meaningful if we measure all $e(i, t, k)$ in the same units of mass (e.g., milligrams per cubic meter) and so admissible transformation means multiply $e(i, t, k)$ by the same constant.

However, although these statements are meaningful in the technical sense, we have to ask if they are meaningful comparisons of pollution level in a practical sense. A unit of mass of CO is far less harmful than a unit of mass of NOX. US Environmental Protection Agency standards based on health effects for a 24 hour period allow many more units of CO than units of NOX since a unit of mass of CO is far less harmful than a unit of mass of NOX. Thus, we might wish to use a weighting factor λ_i that measures the effect of the i^{th} pollutant and then use a weighted sum

$$\sum_{i=1}^{n} \lambda_i e(i, t, k). \tag{10}$$

Such a weighted sum, sometimes known as *pindex*, has been used as a combined pollution index. In early uses of this measure in the San Francisco Bay Area (Bay Area Pollution Control District, 1968, Sauter and Chilton, 1970, and elsewhere), λ_i is the reciprocal of the amount $\tau(i)$ of emissions of pollutant i in a given period of time needed to reach a certain danger level, otherwise called the *tolerance factor*. The reciprocal is called the *severity factor*. Using this version of pindex, Walther (1972) argues that transportation is still the largest source of pollution, but now accounting for less than 50 per cent. Stationary sources fall to fourth place. CO drops to the bottom of the list of pollutants, accounting for just over 2 per cent of the total. Again, these conclusions are meaningful if we use the same units of mass in each case. With these weighting factors $\lambda_i = 1/\tau(i)$, although comparisons using pindex are meaningful in our technical sense, the index does not seem to give meaningful numbers in any real sense. For reaching 100 per cent of the danger level in one pollutant would give the same pindex value as reaching 20 per cent of the danger level on each of five pollutants. In conclusion, we should stress again that there is a distinction between meaningfulness in the technical sense and meaningfulness in other senses.

The *severity tonnage* of pollutant i due to a given source is actual tonnage times the severity factor $1/\tau(i)$. In early air pollution measurement literature, severity tonnage was considered a measure of how severe pollution due to a source was. Data from Walther (1972) suggests the following. It is an interesting exercise to decide which of these conclusions are meaningful in either the technical sense or the practical sense. (i) HC emissions are more severe (have greater severity tonnage) than NOX emissions; (ii). effects of HC emissions from transportation are more severe than those of HC emissions from industry (and the same for NOX); (iii). effects of HC emissions from transportation are more severe than those of CO emissions from industry; (iv). effects of HC emissions from transportation are more than 20 times as severe as effects of CO emissions from transportation; (v). the total effect of HC emissions due to all sources is more than 8 times as severe as the total effect of NOX emissions due to all sources.

6 Evaluation of Alternative HIV Treatments

How do we evaluate alternative possible treatment plans or interventions for a given disease? One common procedure is the following. A number of treatments are compared on different criteria/benchmarks. Their scores on each criterion are normalized relative to the score of one of the treatments. The normalized scores of a treatment are combined by some averaging procedure and average scores are compared. If the averaging is the arithmetic mean, then the statement "one treatment has a higher arithmetic mean normalized score than another treatment" is meaningless: The treatment to which scores are normalized can determine which has the higher arithmetic mean. Similar methods are used in comparing performance of alternative computer systems or other types of machinery.

To illustrate, consider a number of treatments/interventions in the case of HIV: universal screening; free condom distribution; abstinence education; male circumcision, etc. Consider a number of criteria/outcomes: CD4 count (a measure of how well your body is fighting off HIV), days without poor appetite, days without profound fatigue, number days hospitalized, etc.

Table 1 shows three treatments I, II, III and five criteria A, B, C, D, E, with the i, j entry giving the score of the ith treatment on the jth criterion. Table 2 shows the score of each treatment normalized relative to treatment I, i.e., by dividing by treatment I's score. Thus, for example, the 1,2 entry is 83/83 = 1, while the 2,2 entry is 70/83 = .84. The arithmetic means of the normalized scores in each row are also shown in Table 2. We conclude that treatment III is best.

However, let us now normalize relative to treatment II, obtaining the normalized scores of Table 3. Based

Table 1: Score of Treatment i on Criterion j

Treatment/Criterion	A	B	C	D	E
I	417	83	66	39,449	772
II	244	70	153	33,527	368
III	134	70	135	66,000	369

Table 2: Normalizing Relative to Treatment I

Treatment/Criterion	A	B	C	D	E	Arithmetic Mean	Geometric Mean
I	1.00	1.00	1.00	1.00	1.00	1.00	1.00
II	.59	.84	2.32	.85	.48	1.01	.86
III	.32	.85	2.05	1.67	.45	1.07	.84

on the arithmetic mean normalized scores of each row shown in Table 3, we now conclude that treatment I is best. So, the conclusion that a given treatment is best by taking arithmetic mean of normalized scores is meaningless in this case.

The numbers in this example are taken from Fleming and Wallace (1986), with data from Heath (1984), and represent actual scores of alternative "treatments" in a computing machine application.

Sometimes, geometric mean is helpful. The geometric mean normalized scores of each row are shown in Tables 2 and 3. Note that in each case, we conclude that treatment I is best. In this situation, it is easy to show that the conclusion that a given treatment has highest geometric mean normalized score is a meaningful conclusion. It is even meaningful to assert something like: A given treatment has geometric mean normalized score 20 per cent higher than another treatment.

Fleming and Wallace give general conditions under which comparing geometric means of normalized scores is meaningful. It is a research area in measurement theory, with a long history and large literature, to determine what averaging procedures make sense in what situations. We return to this topic, and in particular the Fleming-Wallace conditions, in Section 9.

Table 3: Normalizing Relative to Treatment II

Treatment/Criterion	A	B	C	D	E	Arithmetic Mean	Geometric Mean
I	1.71	1.19	.43	1.18	2.10	1.32	1.17
II	1.00	1.00	1.00	1.00	1.00	1.00	1.00
III	.55	1.00	1.88	1.97	1.08	1.07	.99

7 Meaningfulness of Conclusions from Statistical Tests

Biostatistics is a key component of epidemiological research. However, biostatisticians know very little about measurement theory. Most have never heard about the theory of meaningfulness or limitations that meaningfulness places on conclusions from statistical tests. For over 50 years, there has been considerable disagreement on the limitations scales of measurement impose on statistical procedures we may apply. The controversy stems from foundational work of Stevens (1946, 1951, 1959, and elsewhere), who developed the classification of scales of measurement we have described here. Stevens provided rules for the use of statistical procedures, concluding that certain statistics are inappropriate at certain levels of measurement. The application of Stevens' ideas to descriptive statistics has been widely accepted. However, their application to inferential statistics has been labeled by some a misconception.

To explore these ideas, suppose P is a population whose distribution we would like to describe. We capture properties of P by finding a descriptive statistic for P or taking a sample S from P and finding a descriptive statistic for S. Our examples suggest that certain descriptive statistics are appropriate only for certain measurement situations. This idea, originally due to Stevens, was popularized by Siegel (1956) in his well-known book *Nonparametric Statistics for the Behavioral Sciences*. Our examples suggest the principle that arithmetic means are "appropriate" statistics for interval scales, medians for ordinal scales. On the other side of the coin, it is argued that it is always appropriate to calculate means, medians, and other descriptive statistics, no matter what the scale of measurement. The well-known statistician Frederic Lord made this argument with a famous example of a group of first year college football players who were upset that the average (arithmetic mean) number on their uniforms was less than that of the more advanced players. He argued that it is meaningful for the first year players to average uniform numbers since "The numbers don't remember where they came from." Marcus-Roberts and Roberts (1987) agree. It is always appropriate to calculate means, medians, ... But, they ask: Is it appropriate to make certain statements using these descriptive statistics?

The rest of this section summarizes the conclusions of Marcus-Roberts and Roberts. They argue that it is usually appropriate to make a statement using descriptive statistics if and only if the statement is meaningful. A statement that is true but meaningless gives information that is an accident of the scale of measurement used, not information that describes the population in some fundamental way. So, it is appropriate to calculate the arithmetic mean of ordinal data. It is just not appropriate to say that the mean of one group is higher than the mean of another group.

Stevens' ideas have come to be applied to inferential statistics – inferences about an unknown population P. They have led to such principles as the following:

- Classical parametric tests (e.g., t-test, Pearson correlation, analysis of variance) are inappropriate for ordinal data. They should be applied only to data that define an interval or ratio scale.
- For ordinal scales, non-parametric tests (e.g., Mann-Whitney U, Kruskal-Wallis, Kendall's tau) can be used.

Not everyone agrees. Thus, there has been controversy.

Marcus-Roberts and Roberts argue that the validity of a statistical test depends on a statistical model. This includes information about the distribution of the population and about the sampling procedure. The validity of the test does not depend on a measurement model. That is concerned with the admissible transformations and scale type. The scale type enters in deciding whether the hypothesis is worth testing at all – is it a meaningful hypothesis? The issue is: If we perform admissible transformations of scale, is the truth or falsity of the hypothesis unchanged?

As an example, suppose we have data on an ordinal scale and we consider the hypothesis that the mean is 0. This is a meaningless hypothesis. Can we test meaningless hypotheses? Marcus-Roberts and Roberts say "yes." But they question what information we get outside of information about the population as measured. To be more precise about this, consider how we test hypothesis H_0 about P. We do this through the following steps:

- Draw a random sample S from P.
- Calculate a test statistic based on S.
- Calculate the probability that the test statistic is what was observed given H_0 is true.
- Accept or reject H_0 on the basis of the test.

Calculation of probability depends on a statistical model, which includes information about the distribution of P and about the sampling procedure. But, validity of the test depends only on the statistical model, not on the measurement model. Thus, you can apply parametric tests to ordinal data, provided the statistical model is satisfied. The model is satisfied if the data is normally distributed. Where does the scale type enter? It enters in determining if the hypothesis is worth testing at all, i.e., if it is meaningful.

For instance, consider data on an ordinal scale and let H_0 be the hypothesis that the mean is 0. The hypothesis is meaningless. But, if the data meets certain distributional requirements such as normality, we can apply a parametric test, such as the t-test, to check if the mean is 0.

Similar analyses can be developed for other kinds of statistical tests for data on other types of scales.

8 Optimization Problems in Epidemiology

The impact of climate change includes potential effects on health of humans, animals, plants, and ecosystems. Some early warning signs of climate change include major heat events such as the 1995 extreme heat event in Chicago that led to 514 heat-related deaths and 3300 excess emergency room admissions and the 2003 heat wave in Europe that led to 35,000 deaths. With anticipated change in climate could come an increase in the number and severity of extreme events, including more heat waves, floods, hurricanes, etc. One response to extreme heat events is the evacuation of the most vulnerable individuals to climate-controlled environments. Here, there are modeling challenges, such as: where to locate the evacuation centers; whom to send where; finding ways to minimize travel times from home to evacuation center; etc. Among the problems arising here is the shortest route problem: Find the shortest route from home to evacuation center. This is an example of an optimization problem, more specifically a combinatorial optimization problem. We shall consider the meaningfulness of conclusions about optimality in such problems.

Consider a network with vertices and edges and numbers on the edges representing some sort of strength or level or weight or length of the connection. The problem is to find the shortest path in the network from vertex x to vertex z, where strength of a path is the sum of the weights of edges in it. This problem occurs widely in practice. In the US, just one agency of the US Department of Transportation in the federal government applies algorithms to solve this problem literally billions of times a year (Goldman, 1981). Consider a simple network with vertices x, y and z and edges from x to y with strength 2, y to z with strength 4, and x to z with strength 15. What is the shortest path from x to z in this network? The shortest path is the path that goes from x to y to z, with a total "length" of 6. The alternative path that goes directly from x to z has total "length" 15. Is the conclusion that x to y to z is the shortest path a meaningful conclusion?

The conclusion is meaningful if the strengths define a ratio scale, as they do if they are distances or times as in the evacuation problem. However, what if they define an interval scale? Consider the admissible transformation $\phi(x) = 3x + 100$. Now the weights change to 106 on the edge from x to y, 112 on the edge from y to z, and 145 on the edge from x to z. We conclude that going directly from x to z is the shortest path. The original conclusion was meaningless.

The shortest path problem can be formulated as a linear programming problem. Thus, the conclusion that A is the solution to a linear programming problem can be meaningless if cost parameters are measured on an interval scale. Note that linear programming is widely used in public health as well as in other areas of application. For example, it is used to determine optimal inventories of medicines, assignments of patients or doctors to clinics, optimization of a treatment facility, amount to invest in preventive treatments, etc.

Another very important practical combinatorial optimization problem is the minimum spanning tree problem. Given a connected, weighted graph or network, we ask for the spanning tree with total sum of strengths or weights as small as possible. (A *spanning tree* is a tree that includes all the vertices of the network.) This problem has applications in the planning of large-scale transportation, communication, and distribution networks. For example, given a network, we seek to find usable roads that allow one to go from any vertex to any other vertex, minimizing the lengths of the roads used. This problem arises in the case of extreme events that leave some roads flooded and when we require routes that emergency vehicles can take. Again, it is natural to ask if the conclusion that a given set of edges defines a minimum spanning tree is meaningful. It is surprising to observe that even if the weights on the edges define only an ordinal scale, then the conclusion is meaningful. This is not a priori obvious. However, it follows from the fact that the well-known algorithm known as Kruskal's algorithm or the greedy algorithm gives a solution. In Kruskal's algorithm (Kruskal, 1956, Papadimitriou and Steiglitz, 1982), we order edges in increasing order of weight and then examine edges in this order, including an edge if it does not form a cycle with edges previously included. We stop when all vertices are included. Since any admissible transformation will not change the order in which edges are examined in this algorithm, the same solution will be produced.

Many practical decision making problems in public health and other fields involve the search for an optimal solution as in the shortest path and minimum spanning tree problems. Little attention is paid to the possibility that the conclusion that a particular solution is optimal may be an accident of the way that things are measured. For the beginnings of the theory of meaningfulness of conclusions in combinatorial optimization, see Mahadev, Pekeč, and Roberts (1998), Pekeč (1996a, 1996b), and Roberts (1990, 1994, 1999).

9 How Should we Average Scores?

We have seen that in some situations, comparing arithmetic means is not a good idea and comparing geometric means is. There are situations where the reverse is true. Can we lay down some guidelines as to when to use what averaging procedure? A brief discussion follows.

Let $a_1, a_2 \ldots, a_n$ be n "scores" or ratings, e.g., scores on criteria for evaluating treatments. Let $u = F(a_1, a_2, \ldots, a_n)$. F is an unknown averaging function, sometimes called a *merging function*, and u is the average or merged score.

Fleming and Wallace (1986) take an axiomatic approach to determining appropriate merging functions. They take the case where the domain and range of F are the positive real numbers and consider the following axioms:

- *Reflexivity*: $F(a, a, \ldots, a) = a$
- *Symmetry*: $F(a_1, a_2, \ldots, a_n) = F(a_{\pi(1)}, a_{\pi(2)}, \ldots, a_{\pi(n)})$ for all permutations π of $\{1, 2, \ldots, n\}$.
- *Multiplicativity*: $F(a_1 b_1, a_2 b_2, \ldots, a_n b_n) = F(a_1, a_2, \ldots a_n) F(b_1, b_2, \ldots, b_n)$

They show that if F satisfies these three axioms, then F is the geometric mean, and conversely. It is fairly simple to understand the first two axioms. Reflexivity says that if all ratings are the same, then their average is the same. Symmetry says that the average is independent of the names or order given to the criteria (which might not be true in some applications). The multiplicative property says that the average of the products of the ratings is the same as the product of the averages of the ratings. Fleming and Wallace motivate this axiom by saying that if a_i measures the relative strength of treatment I to treatment II on criterion i, and b_i the relative strength of treatment II to treatment III on criterion i, then $a_i b_i$ measures the relative strength of treatment I to treatment III on criterion i. If "strength" is speed, as in the Fleming-Wallace applications, then there is some justification for this conclusion and, moreover, to the conclusion that the average over criteria of the relative strength of treatment I to treatment III is the product of the average of the relative strength over all criteria of treatment I to treatment II times the average of the relative strength over all criteria of treatment II to treatment III. However, more generally, it is harder to be sure that Multiplicativity is a desired property of an averaging procedure.

An alternative approach uses functional equations and is based on either assumptions about scale type of some of the scales or about meaningfulness of some statements using the scales. Consider an unknown function $u = F(a_1, a_2, \ldots, a_n)$. We will use an idea due to Luce (1959) that he once called the *Principle* *of Theory Construction*: If you know the scale types of the a_i and the scale type of u and you assume that an admissible transformation of each of the a_i leads to an admissible transformation of u, you can derive the form of F. (We will disregard some of the restrictions on applicability of this principle, including those given by Luce, 1962, 1964, 1990.)

To illustrate the ideas, let us take a simple case where $n = 1$, $a = a_1$ is a ratio scale, and u is a ratio scale. An admissible transformation of scale in both cases is multiplication by a positive constant. By the Principle of Theory Construction, multiplying the independent variable a by a positive constant α leads to multiplying the dependent variable by a positive constant A that depends on α. This leads to the functional equation

$$F(\alpha a) = A(\alpha)F(a), A(\alpha) > 0. \tag{11}$$

By solving this equation, Luce (1959) proves that if the averaging function F is continuous and a takes on all positive real values and F takes on positive real values, then

$$F(a) = ca^k.$$

Thus, if the independent and dependent variables are ratio scales, the only possible way to relate them is by a power law.

This result is very general. In early writings of Luce, it was interpreted as limiting in very strict ways the "possible scientific laws" in all disciplines. For example, other examples of power laws are given as follows. One is

$$V = (4/3)\pi r^3,$$

where V is volume and r is radius (both ratio scales). A second is Newton's Law of Gravitation:

$$F = G(mm^*/r^2),$$

where F is the force of attraction, G is a gravitational constant, m, m^* are fixed masses of bodies being attracted, and r is distance between them (everything being a ratio scale). A third is Ohm's Law: Under fixed resistance, voltage is proportional to current (voltage and current being ratio scales).

To illustrate the ideas when the number of independent variables (ratings being averaged) is larger than 1, suppose that $a_1, a_2, \ldots, a_n$ are independent ratio scales and u is a ratio scale. Let F be a merging function defined on all n-tuples of positive real numbers and outputting a positive real. By the Principle of Theory Construction,

$$F(a_1, a_2, \ldots, a_n) = u \leftrightarrow F(\alpha_1 a_1, \alpha_2 a_2, \ldots, \alpha_n a_n) = \alpha u,$$

where $\alpha_1 > 0, \alpha_2 > 0, \ldots, \alpha_n > 0, \alpha > 0$, and α depends on $\alpha_1, \alpha_2, \ldots, \alpha_n$. Thus we get the functional equation:

$$F(\alpha_1 a_1, \alpha_2 a_2, \ldots, \alpha_n a_n) = A(\alpha_1, \alpha_2, \ldots, \alpha_n) F(a_1, a_2, \ldots, a_n), A(\alpha_1, \alpha_2, \ldots, \alpha_n) > 0. \tag{12}$$

Luce (1964) shows that if F is continuous and satisfies Equation (12), then

$$F(a_1, a_2, \ldots, a_n) = \lambda a_1^{c_1} a_2^{c_2} \ldots a_n^{c_n} \tag{13}$$

for constants $\lambda > 0, c_1, c_2, \ldots, c_n$. Aczél and Roberts (1989) show that if, in addition, F satisfies reflexivity and symmetry, then $\lambda = 1$ and $c_1 = c_2 = \ldots = c_n = 1/n$, so F is the geometric mean.

There are also situations where one can show that the merging function F is the arithmetic mean. Consider for example the case where $a_1, a_2, \ldots, a_n$ are interval scales with the same unit and independent zero points and u is an interval scale. Then the Principal of Theory Construction gives the functional equation:

$$F(\alpha a_1 + \beta_1, \alpha a_2 + \beta_2, \ldots, \alpha a_n + \beta_n) = A(\alpha, \beta_1, \beta_2, \ldots, \beta_n) F(a_1, a_2, \ldots, a_n) + B(\alpha, \beta_1, \beta_2, \ldots, \beta_n),$$

where

$$A(\alpha, \beta_1, \beta_2, \ldots, \beta_n) > 0.$$

Even without a continuity assumption, Aczél, Roberts, Rosenbaum (1986) show that in this case,

$$F(a_1, a_2, \ldots, a_n) = \sum_{i=1}^{n} \lambda_i a_i + b,$$

where $\lambda_1, \lambda_2, \ldots, \lambda_n, b$ are arbitrary constants. Aczél and Roberts (1989) show that if, in addition, F satisfies reflexivity, then

$$\sum_{i=1}^{n} \lambda_i = 1, b = 0.$$

If in addition F satisfies reflexivity and symmetry, then they show that $\lambda_i = 1/n$ for all i, and $b = 0$, i.e., F is the arithmetic mean.

Still another approach to determining the form of an appropriate merging function is to replace assumptions of scale type with assumptions that certain statements using scales are meaningful. While it is often reasonable to assume that you know the scale type of the independent variables $a_1, a_2, \ldots, a_n$, it is not so often reasonable to assume that you know the scale type of the dependent variable u. However, it turns out that one can replace the assumption about the scale type of u with an assumption that a certain statement involving u is meaningful. To return to the case where the a_i are independent ratio scales, instead of assuming that u is a ratio scale, let us assume that the statement

$$F(a_1, a_2, \ldots, a_n) = kF(b_1, b_2, \ldots, b_n)$$

is meaningful for all $a_1, a_2, \ldots, a_n, b_1, b_2, \ldots, b_n$ and $k > 0$. Then we get the same result as before: Roberts and Rosenbaum (1986) prove that under these hypotheses and continuity, F satisfies Equation (13). Moreover, if in addition F satisfies reflexivity and symmetry, then F is the geometric mean. For a variety of related results, see Roberts and Rosenbaum (1986).

10 Behavioral Responses to Health Events

Governments are making detailed plans for how to respond to future health threats such as pandemic influenza, H1N1 virus, a bioterrorist attack with the smallpox virus, etc. A major unknown in planning for future disease outbreaks is how people will respond. Will they follow instructions to stay home? Will critical personnel report to work or take care of their families? Will instructions for immunization be followed? Mathematical models are increasingly used to help plan for health events or to develop responses to them. They have been especially important in planning responses to such recent events as Foot and Mouth Disease in Britain and SARS. Models in epidemiology typically omit behavioral responses. These are hard to quantify and hard to measure. This leads to challenges for behavioral scientists and for epidemiological modelers who want to work with them.

In building behavioral responses into epidemiological models, we can can learn some things from the study of responses to various disasters such as earthquakes, hurricanes, and fires. Many behavioral responses need to be addressed. "Compliance" with things like quarantine instructions is one example. How do we measure "compliance?" In particular, this includes such factors as "resistance" to instructions, willingness to seek or receive treatment, credibility of government, trust of decision makers. Other things that need to be made precise and measured include movement, rumor, perception of risk, person-to-person interaction, motivation, social stigmata (such as discrimination against certain social groups), panic, and peer pressure. There is a challenge to measurement theory here: How do we measure some of these factors? How do we bring them into mathematical models? What statements using the new scales of measurement are meaningful? Some of the issues are discussed in McKenzie and Roberts (2003), which summarizes a workshop aimed at modeling social responses to bioterrorism involving infectious agents.

There is much more analysis of a similar nature in the field of epidemiology that can be done with the principles of measurement theory. The issues involved present challenges both for theory and for application.

References

[1] Aczél, J., & Roberts, F.S. (1989). On the possible merging functions. *Mathematical Social Sciences, 17*, 205-243.

[2] Aczél, J., Roberts, F.S., & Rosenbaum, Z. (1986). On scientific laws without dimensional constants. *J. Math. Anal. & Applic, 119*, 389-416.

[3] Bay Area Pollution Control District (1968). Combined pollutant indexes for the San Francisco Bay Area. Information Bulletin 10-68, San Francisco, CA.

[4] Fleming, P.J., & Wallace, J.J. (1986). How not to lie with statistics: The correct way to summarize benchmark results. *Communications of the ACM, 29*, 218-221.

[5] Goldman, A.J. (1981). *Discrete mathematics in government.* Lecture presented at SIAM Symposium on Applications of Discrete Mathematics, Troy, NY, June 1981.

[6] Heath, J.L. (1984). Re-evaluation of RISC I. *Comput. Archit. News, 12*, 3-10.

[7] Helmholtz, H.V. (1887). Zählen und messen. *Philosophische Aufsätze.* (17-52). Leipzig: Fues's Verlag. (C.L. Bryan, transl, Counting and measuring. Van Nostrand, Princeton, New Jersey, 1930.)

[8] Knapp, T.R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing Research, 39*, 121-123.

[9] Krantz, D.H., Luce, R.D., Suppes, P, & Tversky, A. (1971). *Foundations of measurement.* (Vol. I). New York: Academic Press.

[10] Kruskal, J.B. (1956) On the shortest spanning tree of a graph and the traveling salesman problem. *Proc. Amer. Math. Soc., 7*, 48-50.

[11] Luce, R.D. (1959). On the possible psychophysical laws. *Psychol. Rev, 66*, 81-95.

[12] Luce, R.D. (1962). Comments on Rozeboom's criticisms of 'On the possible psychophysical laws'," *Psychol. Rev., 69*, 548-555.

[13] Luce, R.D. (1964). A generalization of a theorem of dimensional analysis, *J. Math. Psychol., 1*, 278-284.

[14] Luce, R.D. (1990). 'On the psychophysical law' revisited: Remarks on cross-modal matching. *Psychol. Rev., 97*, 66-77.

[15] Luce, R.D., Krantz, D.H., Suppes, P., & Tversky, A. (1990). *Foundations of measurement.* (Vol. III). New York: Academic Press.

[16] Mahadev, N.V.R., Pekeč, A., & Roberts, F.S. (1998) On the meaningfulness of optimal solutions to scheduling problems: Can an optimal solution be non-optimal? *Operations Research, 46 suppl*, S120-S134.

[17] Marcus-Roberts, H.M., & Roberts, F.S. (1987) Meaningless statistics, *J. Educ. & Behav. Statist., 12*, 383-394.

[18] McKenzie, E., & Roberts, F.S. (2003). Modeling social responses to bioterrorism involving infectious agents. Technical Report, DIMACS Center, Rutgers University, Piscataway, NJ, July 24. (Available at http://dimacs.rutgers.edu/Workshops/Modeling/.)

[19] Papadimitriou, C.H., & Steiglitz, K. (1982). *Combinatorial optimization: Algorithms and complexity*, Englewood Cliffs, NJ: Prentice-Hall

[20] Pekeč, A.. (1996a). *Limitations on conclusions from combinatorial optimization*, Ph.D. Thesis, Department of Mathematics, Rutgers University.

[21] Pekeč, A. (1996b). Scalings in linear programming: Necessary and sufficient conditions for invariance. Center for Basic Research in Computer Science (BRICS), technical report RS-96-50.

[22] Pfanzagl, J. (1968). *Theory of measurement.* New York: Wiley.

[23] Roberts, F.S. (1972). Building an energy demand signed digraph I: Choosing the nodes. Rept. $927/1 - NSF$. April. Santa Monica, CA: The RAND Corporation.

[24] Roberts, F.S. (1973) Building and analyzing an energy demand signed digraph," *Envir. & Plan., 5*, 199-221.

[25] Roberts, F.S. (1979). Measurement theory, with applications to decisionmaking, utility, and the social sciences, Reading, MA: Addison-Wesley. Digital Reprinting (2009). Cambridge, UK: Cambridge University Press.

[26] Roberts, F.S. (1990). Meaningfulness of conclusions from combinatorial optimization. *Discr. Appl. Math., 29*, 221-241.

[27] Roberts, F.S. (1994). Limitations on conclusions using scales of measurement. In S.M. Pollock, M.H. Rothkopf, & A. Barnett (Eds.), *Operations research and the public sector*, Vol. 6 in *Handbooks in operations research and management science.* (621-671). Amsterdam: North-Holland.

[28] Roberts, F.S. (1999). Meaningless Statements. In *Contemporary trends in discrete mathematics*, DIMACS Series, (Vol. 49, 257-274). Providence, RI: American Mathematical Society.

[29] Roberts, F.S., & Rosenbaum, Z. (1986). Scale type, meaningfulness, and the possible psychophysical laws. *Math. Soc. Sci., 12*, 77-95.

[30] Sauter, G.D., & Chilton, E.G. (Eds.) (1970). *Air improvement recommendations for the San Francisco Bay Area. The Stanford-Ames NASA/ASEE Summer Faculty Systems Design Workshop, Final Report.* October. Stanford CA: Stanford University School of Engineering. Published under NASA Contract NGR-05-020-409

[31] Siegel, S. (1956) *Nonparametric statistics for the behavioral sciences*, New York: McGraw-Hill.

[32] Stevens, S.S. (1946). On the theory of scales of measurement. *Science 103*, 677-680.

[33] Stevens, S.S., (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), *Handbook of experimental psychology* (1-49). New York: Wiley.

[34] Stevens, S.S. (1959). Measurement, psychophysics, and utility. In C.W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories*, (18-63). New York: Wiley.

[35] Stevens, S.S. (1968). Ratio scales of opinion. In D.K. Whitla (Ed.), *Handbook of Measurement and assessment in behavioral sciences*, Reading, MA: Addison-Wesley.

[36] Suppes, P. (1959). Measurement, empirical meaningfulness and three-valued logic. In C.W. Churchman & P. Ratoosh (Eds.), *Measurement: Definitions and theories* (129-143). New York: Wiley.

[37] Suppes, P. (1979). Replies. In R.J. Bogdan (Ed.), *Patrick Suppes* (207-232). Dordrecht, Holland: Reidel.

[38] Suppes, P., Krantz, D.H., Luce, R.D., & Tversky, A. (1989) *Foundations of measurement.* (Vol. II). New York: Academic Press.

[39] Suppes, P., & Zinnes, J. (1963) Basic measurement theory. In R.D. Luce, R.R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, (Vol. 1, 1-76). New York: Wiley.

[40] Thomas, H. (1985). Measurement structures and statistics. In S. Kotz & N.L. Johnson (Eds.), *Encyclopedia of statistical sciences*, (Vol. 5, 381-386). New York: Wiley.

[41] Walther, E.G. (1972). A rating of the major air pollutants and their sources by effect. *J. Air Poll. Control Assoc., 22*, 352-355.