Published entries to the three competitions on “Tricky Stats” in *The Psychologist*

Author’s manuscript
Published entry (within announced maximum of 250 words) to competition on “Tricky Stats” (no. 1) on “confounds”, *The Psychologist* July 1994, 7(10), 437.

Confounding

A factor that varies systematically with an independent variable is liable to create difficulty in the interpretation of an experiment or indeed of observations including an hypothesised predictor. This is the problem of the confounding of experiments (there is no noun "a confound") and of collinearity between predictors in multivariate data.

At the extreme, when two variables (suitably linearised) show a very high correlation, it is not possible to distinguish their effects in the same direction: such a completely confounded experiment is "instantly dead."

However, when the relationship between the independent variable and a confounding variable is less than perfect, then some useful information may be extractable from the experiment or observations. That is, partial confounding is not instant death. For example, if the predicted effect of the confounding variable can be separated out statistically, some evidence of the effect of the independent variable may be found. Nevertheless, the feasibility of such a rescue operation cannot be relied on (covariance analysis is not always valid).

Other 'rescue operations' are occasionally feasible. For example, an experiment (or a multivariate analysis) might be designed where a completely confounding variable is known for sure to have an effect opposite to that expected of the independent variable. Then, if the expected effect is observed, the evidence is that the independent variable has overridden the confounding variable. For instance (in the example given), the fact that slower people are made the speedier ones by the manipulation supports the hypothesis that it should increase speed. However, such evidence is indirect and depends on sound prior knowledge - or sheer luck. Therefore, it is deplorable to "discover" such confounding and designs relying on it should be replaced by direct control.

These are <u>not</u> statistical issues, even though statistical manipulation can sometimes get around mild confounding (see above). These issues are not specific either to experimental design or to any other form of empirical investigation. The same problems can afflict clinical or historical interpretation or literary or philosophical argument.

David Booth 12.7.94

Author's manuscript

TRICKY STATS competition No. 2 (The Psychologist, August 1994)

Answers to questions a), b) and c) from David Booth (UoB), published in December 1994, *The Psychologist* 7(12), 533 (text quoted only).

DOG WAGS TAIL
Big News About P Values

The statistical significance of an observation is NOT the chance that it could have occurred at random. The level of significance (e.g., $P < 0.05$) is a measure of how good the evidence is for the hypothesised effect. That is, the P value represents how far AWAY the observation is from NO effect.

Hence, although any 5% of a random distribution has a probability of 0.05 (even the 5% around the mean), only the extremes (tails) of the distribution signify evidence against the null hypothesis represented by the mean.

What the far-from-null observation is evidence <u>for</u>, though, is not a statistical matter but depends on the design of the test for an effect. What in those conditions could have produced such an extreme? An observation at the mean of a random distribution cannot distinguish among the effects that are not operative(!): that is why the truth of a particular null hypothesis can never be tested or supported.

The appropriate P value to adopt as the criterion of sufficient reliability is also a scientific issue [not a statistical one]. What follows from accepting the evidence?

If the roughest of estimates of an effect is what is needed, then it may be legitimate to adopt a criterion of 10% or even 20% of the tail in the hypothesised direction.

On the other hand, there may be very great theoretical or practical consequences to accepting the observation as evidence for the effect it was designed to test. The best strategy would then be to carry out replicate tests but that may not be feasible. In those circumstances, it might be wise to raise the criterion [of (genuine) 'significance'] to 1 in 100 or even perhaps to 1 in 1000.

D.A.B.
1.8.94

Author's manuscript
Published entry by David Booth to "**Tricky Stats**" competition **III. Levels of measurement**. *The Psychologist* (1995) *8(5),* 196-197.

Scales

Contrary to S.S. Stevens, there are not four types of scale, nominal, ordinal, interval and ratio. The distinction he was referring to (in direly confused terms!) is in fact between listing, ranking and looking for continua, which are normally convertible to scales, i.e. ratio measurement.

[Before going further, note that a FORMAT for expressing quantitative judgments, contrary t o common usage, is NOT a scale in the sense of achieving ratio measurement of any psychological quantity. Indeed, many rating formats, such as "Likert scales", are not adequately designed for a ratin gitem to yield a score that is an undistorted ratio measurement. (The scales that can be obtained from attitude ratings, using Likert-style layouts or better ones, are psychometric constructs from many people's rating scores in response to several question items, as Likert indeed did.)]

An unordered set of labels is simply a LIST. The number of items in the list is a measure of the length of the list, size of the pile etc. Yet using the number series to label items in the list in no sense to put the items on any scale (of length or size, for example), even a "nominal" one. The operation of counting does NOT assign a value of "6" to the item that happens to come 6th in a particular count.

A set of labels that puts items in a sequence is simply a RANK order. The ranks, 1, 2, 3, etc., are likely to represent some measurable continuum but in itself a rank (like 3rd) carries no information as to the interval or ratio between itself and any other rank (like 2nd or 1st). How the ranks are used in data analysis depends entirely on the mathematical assumptions of the analysis, not on what the ranks represent in reality. Thus they may be used as ranks in statistical analyses based on the arithmetic of ranking. If the frequencies of occurrence of different ranks lie close enough to the normal distribution, then variance analyses can be used (ranking data do not require "non-parametric" statistics necessarily). If enough data are available, then underlying intervals and ratios can be estimated. This can provide a scale from a single response format made up of an ordinal sequence of anchor points, e.g. like extremely, like strongly, like, like slightly, neither like nor dislike, etc. (usually scored 9, 8, 7, etc. and commonly misnamed the Hedonic "Scale"). Thurstone measured the distances between these categories in one set of ratings shortly after these categories were introduced by Peryam and Pilgrim. His procedure for estimating psychological distances between responses to ordinal anchors is known as "Thurstone scaling."

When a rating format has only two anchor points and the responses are (to reduce known distortions of measurement) spread fairly evenly over it, without getting very near the ends, then they can be presumed to measure some SCALE in the rater's use of those two anchor categories. (Use of) this format has traditionally been known as a "category scaling" and S.S. Stevens tried to contrast category scoring with use of numerical scoring under instructions to put subjective quantities in ratio (so-called "magnitude estimation"). However, minimally biased two-anchor rating scores can be used by the investigator to estimate the scale used by the assessor. This can be full measurement (the highest "order"), based on ratios as well as intervals: one anchor should always be scored zero, making the distance to the other anchor an arbitrary value. Psychological distances between the two anchor points and all the ratings can be measured in "absolute" units of acuity, given distributions of responses that satisfy the assumptions of the acuity-scaling arithmetic, e.g. equal response variances at all tested stimulus levels.

D.A.B. 15.11.94

Comment to the Editor, The Psychologist, BPS

(based on the competition entry which kept within the limit of 250 words)

'Tricky Stats' No. 3

Please, order an interval for rationality about psychological measurement!

There is no practical distinction between interval and ratio scales in psychology or any other empirical discipline. Indeed, the concept of a 'real' zero is bogus as a requirement for m easuring in equal ratios.

A psychological zero generally is readily available: it may be a norm, an ideal or an indifference point, for example. A zero quantity is not some absolute value, like 0^o Kelvin. It is a useful origin, like 0^o Celsius or 32^o Fahrenheit: -4^oF is three times further below freezing than is 20^oF. What is that 'real' zero for length or time, the paradigms of quantity? The Big Bang? Where and when was that when I want to measure the real distance or duration of the trip from home to work?

[The notion of nominal, ordinal, interval and ratio scales was introduced by S.S. Stevens. It pointed towards some of the distinctions made properly later by theorists of fundamental measurement. The perpetuation of this terminology and its linkage to the choice of statistics and, worst of all, to denial of the reality of quantitative psychology are even more harmful than Stevens's subsequent advocacy of numerical rating under ratio instructions (which means never scoring zero!) as the procedure for directly estimating subjective magnitu des.]

The basic logic of measurement, briefly and simply, is that the data must be categorisable and these categories can be ordered in sequence. If categories can be found that are equi-distant from each other in a sequence, then the highest order of measurement has been attained. When analysed correctly, most psychological data have been found to be fully quantitative.

Like any science, psychology deals with those three main types of data: categories, ranks and quantities.

Categorical data can be quantified and statistically evaluated as the frequencies with which each category is filled - for instance, how many people tick "Yes" rather than "No" or each of 7 boxes from "Strongly disagree" to "Strongly agree", or indeed give a score above a numerical criterion or below it. Which category an individual's behaviour falls into is sometimes called a nominal datum, because the name of the category applies to that person or animal. That name can be a number, like on a football shirt. It is misleading to talk about nominal scales because even a number label does not measure anything. A measure would be something like how many people across the league are labelled by that number or by numbers in a set that falls into some category (like double figures).

Shirt numbers do not provide ranks either, unless perhaps 1 is reserved for the position at the back of a football side and the highest numbers are assigned to the forwards. Rather, a first-rank player scores or saves more goals than most or gets a record transfer fee.

In psychology we sometimes rank-order people or collect data that can be converted into rankings. There are exact statistical tests for differences between categories of people in the distributions of ranks (e.g. Mann-Whitney, Wilcoxon, Kruskal-Wallis). These tests are useful if the raw data are just one person per rank or for some other reason cannot be transformed to meet the assumption of normal distributions that is made in analysis of variance or correlational statistics.

However, we often rank what people do, not the people themselves. For example, from the plain meaning of the words, the answer "Strongly disagree", "Disagree" or "Slightly

disagree", etc., can be assigned a response rank such as -3, -2, -1, etc., respectively. Unless we have shown previously that these numbers represent quantities (e.g. +3 is three times as far from "Neither agree nor disagree" as +1 is), they are merely an ordering of the response categories, namely ordinal data. Unlike plain ranks, however, such data can be near-enough normally distributed and hence subjected to factor analysis, multiple regression, ANOVA and so on. That is, "parametric" statistics are <u>not</u> precluded by a merely ordinal level of measurement. Those analyses do not depend on parameterisation of the investigation; they depend on normalcy of distribution of the numbers..

Thus, neither those multiple-category rating formats nor the ranked scores, "Strongly a gree" = 3, etc., should be confused with the attitude scales that Likert and others constructed by statistical analysis of such answers from substantial samples of respondents to a variety of such questions. A respondent's score on one of these multi-item factors or sub-scales is a measure of how strongly that person holds that attitude: it is a psychological quantity which can be represented by a real number, including zero in principle (although total neutrality of attitude might never be observed).

David Booth
30.1.95

Notes to Editor

1. The paragraph in square brackets can be omitted or moved to the end. If the paragraph is retained, its last 3 lines can be deleted, ending (after “statistics”) "...are deplorable."

2. This piece's jokey title probably deserves to survive even less than my previous two did in your kind publication of my submissions.

D.A.B.