Are you behind the times?

Kate Grey

David Clark-Carter *introduces the special issue on statistics.*

STATISTICS, like any branch of human knowledge, is open to debate and subject to change. However, many psychologists have been happy to stick to a set of techniques that they learned in their first year as undergraduates and that have been around since the mid-1930s. Statisticians have continued to explore those techniques and create new ones, and the availability of ever-faster computers has meant that aspects of the techniques can be tested in ways that were never available to those who originated them. In 1900 when Pearson devised the χ^2 test, which is often considered the first 'modern' statistical test, computer and calculator referred either to a pretty rudimentary mechanical device or even just a person with a sharp pencil and a large piece of paper. Things have moved on. Nonetheless, numerous attempts to persuade psychologists (and those in many other disciplines) of the need to take account of the new findings and approaches have fallen on stony ground.

Now that first-year undergraduates have access to computer facilities that could only be dreamed of even 10 years ago, it is possible to employ quite advanced techniques at the touch of a button. This makes it all the more important that we understand when specific tests are appropriate, what their limitations are, how to interpret what they are telling us, and how to present the information to others. Without this understanding we are in ever-greater danger of GIGO – garbage in, garbage out – and now we can achieve it to as many decimal places as the computer will let us.

If that were not reason enough, there are other pressures that mean psychologists can no longer ignore these issues. The APA convened a working party to discuss these matters, and, although not everyone is happy about how many of its recommendations have been implemented, it has had an impact on the guidelines in the latest edition of the *APA Publication Manual*. A second influence for change is ethics committees, in particular those that are linked to work with patients. They now expect justification for statistical techniques and choice of sample sizes.

The following articles form a somewhat ad hoc set, but one which reflects the interests and concerns of each of the authors. They are designed to extend the range of issues which you are aware of in this area and, even if you are not going to use the techniques described yourself, to enable you to be in a better position to evaluate research that employs them.

BIG with

Pasco Fearon *describes how computers can come to the rescue with resampling methods.*

AS the person responsible for teaching statistics on the UCL Clinical Psychology course, I often have the dubious privilege of acting as troubleshooter for clinical students during their training (and staff, let's be honest). One of the most common questions is when to worry about small samples or non-normal distributions.

A typical scenario runs something like this: A trainee comes to see me and says, 'I've got these three groups and I want to know whether the differences between them in the proportion of cases with depression is significant or not. But when I ask SPSS to do a chi-squared test, it warns me that more than 50 per cent of my cells have an expected count of less than 5. What should I do?' Until fairly recently, I might have said, 'Oh, just do it, but *sound cautious.*'

A very similar response can be triggered by a trainee saying, 'I want to do a *t* test/ANOVA/regression, but I don't think my dependent variable is very normally distributed – it certainly doesn't look like a bell curve.'

These two related problems are critical issues in everyday data analysis. Most of us have some idea that statistical tests cannot be trusted when the sample size is small, or when the data are not normally distributed – but knowing there is a problem is still a long way from knowing what to do about it.

Assumptions of statistical testing

To understand this problem, and to see how resampling methods can help, we need to briefly remind ourselves about the essential mechanics of statistical hypothesis testing. Imagine we carry out assessments of autobiographical memory with 50 people suffering from depression, and compare the result with a group of 50 people who are not. The measurements that we have taken are proxies, or estimates, for quantities that describe whole populations (e.g. all people with depression). This short cut is convenient and sensible, but it is subject to error. Faced with a dish in a restaurant that you haven't had before, you might very reasonably have a nibble to see if you like it, and if you do, plough in and eat the lot. This initial nibble gives you an estimate of what the whole thing tastes like. However, unless the dish is completely homogeneous, you can't be sure that the nibble you had was just a particularly nice bit of what is actually a rather rancid concoction.

This is exactly the kind of problem that statistical methods are designed to help us with. They consist of a set of methods for estimating population quantities from samples, and, most importantly, they give us a gauge of how much the sampling process itself is affecting how variable our estimates are likely to be. Going back to the example on depression, let's say your measurements gave you a mean score for autobiographical memory of 31.2 for the depression group and 45.6 for your control group. You might confidently declare that autobiographical memory is poorer in depression if you knew that the sampling process introduces only a small amount of variation in your estimates (say 5 points up

problems
small samples

or down). However, if you learnt that the sampling process could commonly introduce enough variability to push either mean up or down by 20 points, you would feel much less confident doing so. In fact, we can never know for certain whether these two scores came from populations with different means, or were just two samples from the same population whose means just happened to be nudged away from each other by the sampling process.

The way that statistical methods typically deal with this problem is to turn it around and ask: 'OK, if we assume these two groups did come from the same population (this is the null hypothesis), how often would the sampling process produce two samples with means that differ by this much?' If we find that sampling from one population could easily produce differences in means of this size or bigger, we are inclined to assume that these samples probably did come from the same population (we accept the null hypothesis). If the probability of getting such differences in mean between samples taken from the same population is low (< .05), we reject the idea, and infer that the samples must have come from different populations. This is the basis of statistical hypothesis testing.

So, getting a grip on the variability that is introduced to estimates as a result of the sampling process is the key to making statistical decisions. The way that traditional statistical methods do this is by developing complicated mathematical equations from first principals, based on assumptions about the nature of the sampling process and the kind of population that the estimates come from. This is where the problems come in. First, the ideas behind these mathematical solutions are hard to understand, and lack a certain amount of intuitive appeal. Second, the assumptions on which these methods are based mean they can only be trusted under certain, quite restricted, circumstances. In particular, most are based on results that only hold true when either the populations follow a normal distribution or, when they don't, the samples that you take are large. If neither of these is the case, the tests are approximate and can sometimes be quite far off the mark.

Resampling methods give us a new way of doing statistics that is much easier to understand, generally makes fewer assumptions, and can be trusted in circumstances where traditional methods cannot. As things stand at the moment in psychology, these methods are underused, although the greater availability of them in the last few years is beginning to change that. To give you a flavour of them, I will describe two important examples – permutation tests and bootstrap resampling.

Permutation tests

Permutation tests (so-called, but really based on combinations) have been around for a long time. Fisher (1925) suggested them many years ago, and Fisher's exact test is an example of one. The logic of the method is really quite different to what we are used to. Rather than making inferences about samples and their relation to populations, permutation tests are based purely on the samples themselves. In his recent book Phillip Good (1999) gives a nice example, based on an apparently true story in which he conducted an experiment on vitamin E supplements and their effects on ageing. He put human cells in six Petri dishes, half with the supplement, half without. When he came to measure their growth, he discovered he had lost the labels that marked which were which. The six Petri dishes had the following growth readings: 121, 118, 110, 34, 12, 22. Good thought to himself: 'If the first three were the treated ones, I've discovered the

TABLE 1 An example of a permutation test – combining secure/insecure attachment scores in different groupings to establish the probability of the actual result being due to chance

'Insecure' Scores	*'Secure' Scores*	*Sum of 'Insecure'*	*Sum of 'Secure'*	*Difference (in descending order)*
43 47 50 58	30 30 34 41	198	135	63
41 47 50 58	**30 30 34 43**	**196**	**137**	**59 (Actual result)**
41 43 50 58	30 30 34 47	192	141	51
34 47 50 58	30 30 41 43	189	144	45
43 41 47 58	30 30 34 50	189	144	45
30 47 50 58	30 34 43 41	185	148	37
34 43 50 58	30 30 41 47	185	148	37
34 43 50 58	30 30 41 47	185	148	37
34 41 50 58	30 30 43 47	183	150	33
34 43 47 58	30 30 41 50	182	151	31
(A further 60 combinations are possible)		…	…	…

fountain of youth, otherwise I have nothing to report.' With the labels lost, there are lots of ways in which the 'vitamin E' labels might have originally been stuck to three of those six dishes. Only a minority of those would suggest an exciting result. This idea is the basis of a permutation test. Take your readings and (a) work out the quantities that you're interested in (e.g. the difference in total growth between the two groups of dishes), then (b) rearrange the labels in every possible way, each time working out the totals and the differences between them, and (c) see how many of those produce values as big as the one you started with (this assumes that you didn't actually lose the labels).

Let's use another example to make things clearer. Say we have scores for state anxiety from two groups of people: those who, according to self-reports, have an insecure attachment relationship with their mother, and those who report a secure relationship. The labels for these cases were not lost, thankfully: the scores for the first group were 41, 47, 50 and 58, and for the second group 30, 30, 34, and 43. The difference between the totals for these two groups is 59.

Following the permutation procedure described above, we now shuffle the labels (secure and insecure) between the scores in all possible ways and find out how many of those produce a difference in total of 59 or more. There are actually 70 different ways that you could reshuffle the labels in this case. In Table 1 I have shown the 10 combinations that produce the largest differences, ranked in descending order.

As you can see, of the 70 ways that you can put the scores together into two groups, only two of them produce a difference as big as 59. So, if there is nothing but chance at play in our data then it seems we have witnessed a relatively rare event. Thus, we can say that the probability of getting the result we did by chance is 2 out of 70 or 2/70 = .029, and we would consider this significant. This is an exact result, which is not based on any approximation, and makes minimal assumptions about the nature of the data. This method is remarkably straightforward, and doesn't require any complicated mathematics. It is also very flexible. For example, you could compute all sorts of other figures from the data, like the mean, the variance, the sum of squares, and so on. Permutation tests are also very useful for analysing single-case data (see Todman & Dugard, 2001).

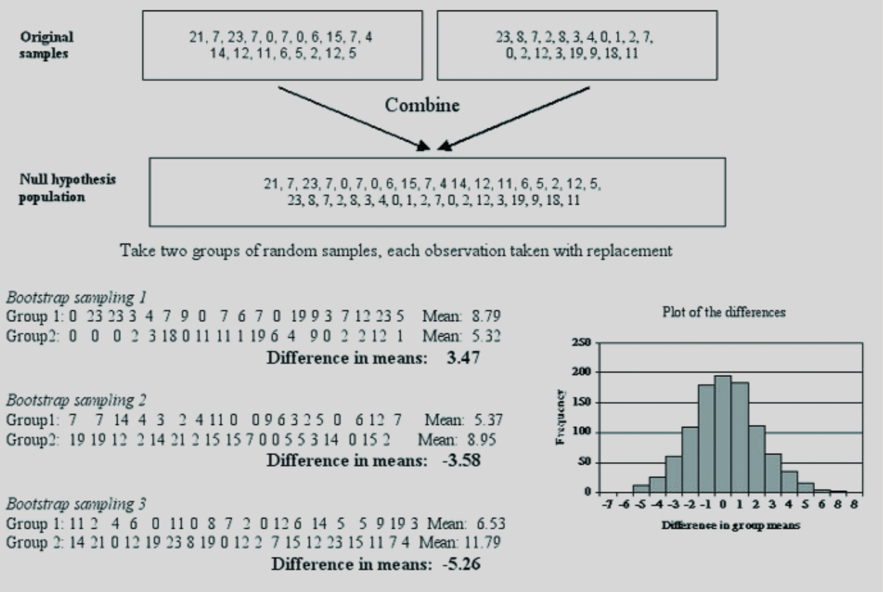
Obviously, permutation tests are laborious because there are so many rearrangements that you need to crunch through, for anything but the smallest of samples. However, computers are great at doing that kind of thing and short cuts have recently been developed that allow the process to be done very efficiently (Mehta, 1990). Some of these have been incorporated into SPSS's 'exact statistics' options. So now when you use SPSS's non-parametric statistics you can always ask for an exact *p* value, rather than an approximate one. An even wider range of tests is available using the program StatXact (Mehta, 1990). Generally speaking, permutation methods are the method of choice when you cannot be certain that the distribution you are sampling from is normal and your sample is relatively small (say, less than 40 cases in each group you have). For more details see Good (1999).

But permutation methods aren't available for some situations, like multiple regression or unbalanced ANOVA designs. For these, we can turn to the bootstrap.

The bootstrap

The bootstrap, another resampling technique, was first introduced in the late 1960s, but was worked out in detail by Efron (1983). The procedure is quite similar to permutation tests in many respects, although the focus is more directly on the sampling process. The basic idea is related to the discussion above – that the key to statistical estimation and testing is working out how much your

FIGURE 1 Bootstrap method for hypothesis testing with two independent groups



Take two groups of random samples, each observation taken with replacement

Bootstrap sampling 1
Group 1: 0 23 23 3 4 7 9 0 7 6 7 0 19 9 3 7 12 23 5 Mean: 8.79
Group2: 0 0 0 2 3 18 0 11 11 1 19 6 4 9 0 2 2 12 1 Mean: 5.32
Difference in means: 3.47

Bootstrap sampling 2
Group1: 7 7 14 4 3 2 4 11 0 0 9 6 3 2 5 0 6 12 7 Mean: 5.37
Group2: 19 19 12 2 14 21 2 15 15 7 0 0 5 5 3 14 0 15 2 Mean: 8.95
Difference in means: -3.58

Bootstrap sampling 3
Group 1: 11 2 4 6 0 11 0 8 7 2 0 12 6 14 5 5 9 19 3 Mean: 6.53
Group 2: 14 21 0 12 19 23 8 19 0 12 2 7 15 12 23 15 11 7 4 Mean: 11.79
Difference in means: -5.26

estimates (say the mean, median, variance, or regression coefficient) are subject to error because of the sampling process. In traditional methods, this variation is calculated using complicated mathematical equations. The bootstrap method is much simpler – and it doesn't assume that your sample is normally distributed.

The bootstrap uses your group of observations itself to *simulate* the sampling process. It does this in a remarkably simple way: take your sample of observations, pretend it is the actual population, and take random samples from it. The samples that you take are done with replacement, which means that each time you take an observation you put it back, before taking the next one. This means that the same observation could appear in your sample more than once. You then calculate the quantity of interest (say, the mean), note it down, and then repeat the whole process again. You do this many times – 1000 perhaps. Sampling repeatedly like this gives you lots of different means, which can be plotted on a histogram to give you some idea about how variable those means are likely to be because of the sampling process. The resultant plot is known as an empirical sampling distribution (rather than a theoretical one, like the normal distribution or the *t* distribution).

Let's work through another example to get a feel for the process in action. We are going to look at differences in the average number of children's errors in a continuous performance test between girls and boys. Remember that hypothesis testing involves a null hypothesis – in this case that the two groups of scores were sampled from the same population. In bootstrapping, we want to construct a population that mimics this null hypothesis and then simulate the sampling process. There are various ways of doing this (see Efron & Tibshirani, 1993), but one simple method is just to combine your two sets of observations, reflecting the fact that the null hypothesis assumes they are both taken from the same population. This then becomes our simulated null hypothesis population. Remember, the logic of this kind of analysis goes like this: when you take two samples from the same population and compare their means, how often could you get a difference as big as the one we actually got? The bootstrap approach mirrors this logic by taking two groups of samples at random from the null hypothesis population (our two samples merged together) and repeats this process many times.

At the top of Figure 1, I have shown the data for boys and girls. Their means were 8.63 and 7.32 respectively (difference = 1.31). To create a simulated null hypothesis population I then combined the two samples. When I repeatedly took two groups of observations from the null hypothesis population (with replacement), I found that 254 out of 1000 produced two groups with a difference in mean as big as if not bigger than, the difference we actually got (1.31). We can produce a histogram of the frequencies of the differences in mean that were produced from the bootstrap sampling: another example of an empirical sampling distribution, this time of the expected distribution of differences between two independently sampled means (assuming the null hypothesis).

Next, we can convert our bootstrap sampling into a probability by dividing 254 by 1000. That gives us a *p* value of .254 (one-tailed). In other words, this bootstrapping simulation suggests that when sampling from the same population – if the null hypothesis is true – you can get a difference in mean between two groups as big as 1.31 25 per cent of the time (p = .254). The whole process is summarised schematically in Figure 1.

Bootstrap methods can be applied to a very wide range of statistical procedures, including all the simpler ones, like the *t* test, or analysis of variance. But it is also able to handle more complex ones, like regression, discriminant analysis, generalised linear models or even structural equation modelling.

The bootstrap is not an exact method, unlike the permutation tests, but nevertheless works well in situations where the normal distribution can't be relied upon. Currently, there are relatively few software packages that do these analyses comprehensively, but that will change in the future. The programs STATA and SAS have implemented bootstrap methods for most circumstances, but currently SPSS uses them only for the non-parametric tests and for non-linear regression.

Time for a rethink about resampling

Computers are changing the way we do statistics, and these changes are leading to methods that are easier to teach and perform well in situations where sample sizes are relatively small and we can't be sure distributions are normal. In these situations we should probably not rely on traditional methods, but use resampling techniques instead. It is quite possible that if you don't, your results will be misleading, and potentially important discoveries will be missed.

■ *Pasco Fearon is in the Department of Psychology, University College London. E-mail: ucjtcrf@ucl.ac.uk.*

References

Efron, B. (1983, May). Computer intensive methods in statistics. *Scientific American*, pp.116–130.

Efron, B. & Tibshirani, R.J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fisher, R.A. (1925). *Statistical methods for research workers*. Edinburgh: Oliver & Boyd.

Good, P. (1999). *Resampling methods: A practical guide to data analysis.* Boston: Birkauser.

Mehta, C.R. (1990). StatXact: A statistical package for exact nonparametric inference. *Journal of Classification, 7*, 111–114.

Todman, J.B. & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests.* Mahwah, NJ: Lawrence Erlbaum.

Effect size

The missing piece in the jigsaw

David Clark-Carter *on why you just can't carry on reporting statistical significance alone.*

JUST when you've got to grips with how to decide whether the result of your research is statistically significant, you find that some people are suggesting that you need to do things differently. The latest edition of the American Psychological Association's *Publication Manual* identifies 'failure to report effect sizes' as one of the 'defects in the design and reporting of research' (APA, 2001, p.5). In addition, the British Psychological Society now has a statement in the 'Notes for contributors' for all its journals that in normal circumstances, effect size should be incorporated. This article explains why statistical significance, on its own, is inadequate as a way of deciding the worth of a piece of research and suggests how we can use a combination of types of evidence to come to our decisions.

By the 1930s some psychologists had already been using a form of statistical analysis similar to today's. However, it wasn't until they had read Fisher (1935) that the process became formalised and psychologists started to use the tests and some of the conventions that we now employ. According to that approach, a null hypothesis is stated – the equivalent of saying that there is no effect, when looking at differences between groups, or no relationship in the case of correlation. Thus, if our study was into the effectiveness of an intervention to reduce truancy compared with a control group who received no intervention, then the null hypothesis is likely to be *The intervention and control groups show the same amount of truancy*. If the study was about the relationship between two variables, for example locus of control (LoC) and willingness to initiate a conversation in a job interview, the null hypothesis might be *There is no relationship between LoC and willingness to initiate a conversation in a job interview*.

The data from the study are collected and analysed and a probability is produced. That probability is answering the question: How likely would be the result of my study if the null hypothesis were true? We then decide whether that probability is sufficiently unlikely that we can reject the null hypothesis; conventionally we decide to reject the null hypothesis if the probability is 0.05 or less. Now, given the nature of that probability, we know that our result could have occurred even if the null hypothesis were true. Therefore, we are risking making an error if we choose to reject the null hypothesis (what has come to be known as a Type I error – rejecting the null hypothesis when it is true).

This procedure can be seen as an odd way round of doing things. We have a notion of what may be the case (that our intervention will be effective or that there is a relationship between LoC and willingness to initiate a conversation). But instead of testing that, we propose a null hypothesis that it won't be the case and then see how likely our result would be if the hypothesis we don't believe in were to be true. The reason for this is that we cannot prove that something is the case, however much evidence we accrue in its favour, whereas we can demonstrate that something is unlikely to be the case if there is little evidence for it (i.e. that the null hypothesis is unlikely to be true) (see Popper, 1974).

Neyman and Pearson (1933) went on to expand the procedure so that our research (or alternative) hypothesis became part of the process. Thus we might propose that Those in the intervention group will truant less than those in the control group or that there is a negative relationship between LoC and willingness to initiate a conversation in an interview (a high score on LoC meaning that a person feels that he or she is controlled by external forces). The introduction of the alternative hypothesis led to the possibility of our making a second type of error (a Type II error) of rejecting the alternative hypothesis when it

Kate Grey

How can statistics show who will speak first?

was true. As Gigerenzer *et al*. (1989) point out, what we now have is a hybrid version of Fisher's and Neyman and Pearson's methods that neither would wholly endorse.

What is often ignored from Neyman and Pearson's contribution is the notion of statistical power: the probability of accepting the alternative hypothesis when it is true. (This probability is related to the probability of making a Type II error through the equation: power = 1 – probability of a Type II error, which also means that probability of a Type II error = 1 – power.) A problem with the approach that Fisher popularised is that failure to achieve statistical significance could be due not just to the null hypothesis being true. It could also be because although the null hypothesis was false, there was insufficient statistical power to achieve a statistically significant result. For example, if a study had only power of 0.3 (and therefore the probability of a Type II error was 0.7) then the likelihood of committing a Type II error is much greater than that of avoiding one. On the other hand, achieving statistical significance could be due to two situations: either the alternative hypothesis is true, or we have such a high level of statistical power that even a trivial result is shown to be statistically significant. In the absence of another piece of the jigsaw we cannot know what statistical significance, or the lack of it, is telling us. That missing piece is effect size.

In Molière's *Le Bourgeois Gentilhomme* a character is surprised to find that he has been speaking prose for 40 years without realising it. As I hope to demonstrate, many of you may have been reporting some effect sizes for years. However, the need is to make this a more conscious and consistent practice.

The advantage of an effect size over an inferential statistic, such as the *t* test, is that it is relatively unaffected by the sample size (although as with any statistic, the larger the sample the closer the effect size will be to that which would have been found if you had had data from the whole population). Thus, an effect size can be used to compare studies that have used different sample sizes, whereas their test statistics, and probabilities, can only sensibly be compared if their sample sizes are the same. Effect sizes take a number of forms and each form has its own variants. Nonetheless, it is possible to describe just three to cover most types of design and data, and even they can be converted into a standard form. Accordingly, I am going to concentrate on the most common forms of effect size that could be reported in the following situations. In most cases the effect size is not the same as the inferential statistic which we use to find statistical significance.

Effect size and *t* tests

In this case we are looking for differences between two conditions where the measure being taken is more than categorical; in other words, as in the truancy intervention example, it is more than simply the number of people falling into a given category.

One of the most popular effect sizes is Cohen's *d* (Cohen, 1988). This is a measure of the difference between the means of the two groups being compared. However, the difference is divided by the standard deviation, which standardises the result so that we are told how many standard deviations the two groups are apart.

The statistical significance trap

Before moving on to other effect sizes, now that you know one it is possible to demonstrate the problem of evaluating the outcome of a piece of research relying solely on statistical significance. First, it is useful to mention the work of someone who spent nearly four decades trying to persuade psychologists and others of the need to report effect sizes and consider statistical power: Jacob Cohen. Cohen (1988) reports work he conducted to measure the effect sizes found by behavioural scientists using various designs and data. For each situation he described what he considered to be a small effect, a medium effect and a large effect. For *d* he said that 0.2 (or just under a quarter of a standard deviation difference between the conditions) is a small effect size, 0.5 (or half a standard deviation) is a medium effect and 0.8 (or over three-quarters of a standard deviation) is a large effect size.

Imagine that two different research groups have devised interventions to reduce truancy rates; each group has devised a different intervention. They each carry out their study and analyse the data employing a *t* test with a one-tailed probability. The result from the first group is that $t = 2.47$, $p = .008$. They conclude that the intervention is effective. The second group's results are $t = 1.5$, $p = .078$. They conclude that their intervention is not effective. However, if we look at the effect sizes, we find that the first group only managed to reduce truancy by an average of half a day in a whole school year, which resulted in an effect size of 0.05, well below what Cohen considers a small effect. On the other hand, the other intervention reduced truancy by an average of seven and a half days, giving an effect size of 0.75, which is what Cohen considers approaching a large effect size.

The reason for these anomalous findings is that the two groups used very different sample sizes and accordingly had very different levels of statistical power. The first study used a sample of 4900 children in each group; the second study employed only eight in each group. Despite the small effect size and because of the large sample size, the first study had power of 0.8, which is the commonly recommended minimum, while, despite the large effect size and because of the small sample size, the second study had statistical power of only 0.4. This means that the risk of committing a Type II error in the second study was 1 – 0.4, or 0.6. In other words, the likelihood of committing a Type II error was 60 per cent. It is possible to find the necessary sample size to achieve power using tables (see Cohen, 1988, or Clark-Carter, 1997) or using software such as G*Power (see Erdfelder *et al*., 1996).

If you doubt whether studies use numbers of participants as large as the first study or as low as the second, then I can assure you that they do. It is worth noting that if a study used as many as 22 participants in each group they may well have failed to achieve a significant result even though the intervention reduced truancy by one school week (a *d* of 0.5: a medium effect size).

Effect size for correlation

This is one example where the standard inferential statistic – Pearson's *r* – can be treated as a measure of effect size, and so it is likely that you have been reporting effect sizes after all. It has the advantage that it is constrained in the values that it can take. We know that a perfect positive correlation cannot be bigger than 1, a perfect negative correlation cannot be larger than –1 and a correlation of zero would suggest no correlation at all.

Effect size for ANOVA

Here there are more than two conditions being compared; for example, participants' intentions of eating a low-fat diet may be compared across people who are in five different stages of a model about how people plan their behaviour. The same effect size can be employed when there is more than one independent variable; for example, a comparison of the physiological responses of three groups of people – those who were socially phobic, those who were anxious and those who were neither anxious nor phobic – on a number of tasks designed to provoke anxiety or phobia, or simply to be physically demanding.

One effect size that can be used in this context is η^2 (eta squared). This is the measure of the proportion of variance which can be accounted for in the dependent variable by differences in the levels of an independent variable. Thus, in the case of participants' intentions to eat a low-fat diet, we might find that differences between the people who were in different stages of the model could account for a certain proportion of the variance in the intention to eat a low-fat diet.

Effect size and multiple regression

Say, for example, that level of depression is being predicted among people who have hepatitis C from measures such as their social functioning and their physical functioning. Here is another example where, often unwittingly, psychologists have been reporting an effect size. The square of the multiple correlation coefficient (R^2) is like η^2, in that it is the proportion of variance in the dependent variable (or variable to be predicted) that can be accounted for by the independent variable(s) (or predictor variable(s)). In fact, if you square Pearson's *r* you have another measure of the proportion of variance in one variable which can be accounted for by the variance in the other. Some people prefer r^2 to *r* as an effect size in correlation because of this interpretation.

Effect size and χ^2

Imagine the data are categorical, for example looking at the types of auditory hallucination in patients with schizophrenia and patients with tinnitus to see whether the hallucinations where predominantly verbal or musical. Here it is likely that a χ^2 test would be employed. One effect size for this situation is *w*, which is actually the same as ϕ (phi) (for a 2x2 contingency table) or Cramér's ϕ (for a larger table). It can be treated like Pearson's *r* in that the closer it is to 1 the stronger the relationship – for example between patient category and type of hallucination.

Effect size and meta-analysis

I mentioned that the different effect sizes can be converted to a common form. This is what is done in meta-analysis. The process may seem puzzling but the principles that underlie many of the tests we use are the same, so it is often possible to make such a conversion. Typically, other effect sizes are converted to either *d* or *r*. It is then possible to create a combined effect size that summarises the results of a number of studies. One of the many benefits of routinely reporting effect size is that the results of a given study can be entered into a meta-analysis more straightforwardly.

Why effect size is an ethical issue

I hope that I have demonstrated that

- on its own, a probability level does not tell us more than that we have (or have not) used a sufficient sample size to achieve a statistically significant result;
- a measure of the magnitude of the effect provides important supplementary information and should be reported routinely; and that
- given its role in statistical power, it makes little sense to pick a sample size on an arbitrary basis.

The choice of an adequate sample for the purposes of a given piece of research is an ethical issue: sampling either too few or too many constitutes a waste of participants' time.

■ *Dr David Clark-Carter is at Staffordshire University.*
E-mail: d.clark-carter@staffs.ac.uk.

References

American Psychological Association (2001). *Publication manual of the American Psychological Association* (5th edn). Washington, DC: Author.

Clark-Carter, D. (1997). *Doing quantitative psychological research: From design to report.* Hove: Psychology Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edn). Hillsdale: NJ: Lawrence Erlbaum.

Erdfelder, E., Faul, F. & Buchner, A. (1996). Gpower: A general power analysis program. *Behavior Research Methods, Instruments, and Computers, 28*, 1–11.

Fisher, R.A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J. & Krüger, L. (1989). *The empire of chance: How probability changed science and everyday life.* Cambridge: Cambridge University Press.

Neyman, J. & Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, Series A, 231*, 289–337.

Popper, K.R. (1974). *Conjectures and refutations: The growth of scientific knowledge* (5th edn). London: Routledge.

Structural equation modelling…

Navigating spaghetti junction

Jonathan Berg/www.bplphoto.co.uk

Obligatory [structural equation models]…diagrammatic spaghetti with largely forgettable and usually unreplicable conclusions. (Barrett, 2002)

Jeremy Miles *and* **Mark Shevlin** *explain why all those boxes and arrows are worth the effort.*

WE'VE all been force fed these spaghetti diagrams, whether it is in journal articles or on overheads at conferences. They look complicated to do, and difficult to interpret. So are they worth the effort?

The chances are you fall into one of two camps on this. As a modern-day psychologist, no doubt you will have developed or used sophisticated and elaborate theories. But perhaps when you want to test these theories, you squeeze them into that familiar old 2x2 ANOVA design. Whilst it is down to disciplines other than psychology to come to firm conclusions about whoever did design the world, we can be fairly sure that they didn't use a 2x2 matrix. Yet here we are doing 21st century psychology tied up in a straitjacket of statistical analysis that was invented in the 1920s.

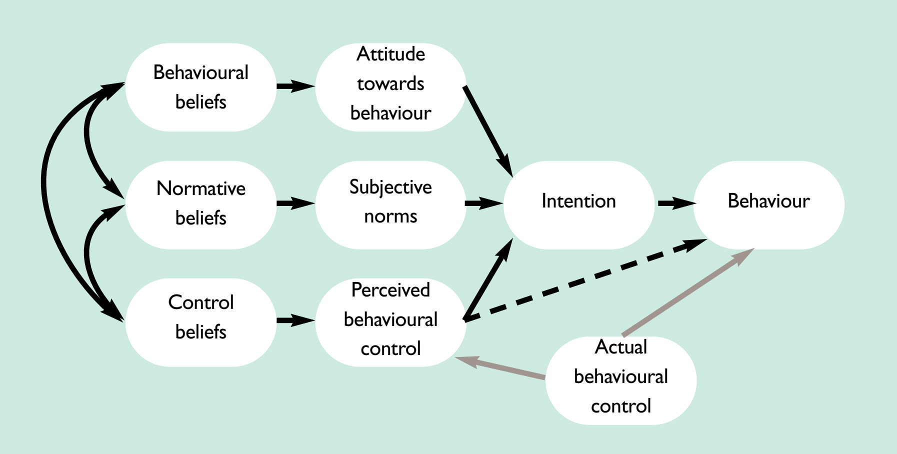
Or perhaps your computer has freed you. Many problems that were considered just too difficult in the past have become possible, and people routinely apply statistical techniques in ways that the people who developed those techniques would never have dreamt of. Structural equation modelling (SEM) sounds like it should be hard, and indeed when structural equation first emerged (in the form of LISREL, by Jöreskog and Sörbom) it was challenging. Many people believe that this is still the case – but they have not actually tried to do it for a long time. If you last tried to do multiple regression in the 1970s, you'd think that was difficult too.

Many of the inferential tests that are commonly carried out can be represented using SEM (e.g. *t* tests, ANOVA, regression). However, there is little point adopting a whole new way of thinking about analysis only to then do what we have always done. The use of computers has led to a whole new way of thinking about problems (bootstrapping, Monte Carlo simulation, multilevel modelling, multidimensional scaling, item response theory… the list could continue). The purpose of thinking about analysis, and hence our theories, in terms of SEM is to think in new ways about data. We can do this in two ways using SEM – we can think in terms of more complex relationships between variables, and we can think in terms of latent variables.

Complex relationships

Psychological theories frequently propose complex relationships between variables. Take Ajzen's theory of planned behaviour (Figure 1). The theory is complex: there are direct effects between variables, indirect effects (such as the effect of subjective norm on behaviour via intention), and mediating variables. In addition, these effects may differ between different groups, and many researchers have proposed that additional components could be added to improve the model (e.g. Fekadu & Kraft, 2001; Heath & Gifford, 2002). SEM provides a means of testing this model – it can tell us if the model could be a plausible one to account for the data. In addition, it provides the flexibility of including additional variables or specifying alternative formulations of the model. The conventional statistical toolkit would be hopelessly inadequate for such a job.

FIGURE 1 Representation of the theory of planned behaviour



Latent variables

The second advantage of thinking in terms of SEM is the ability to use latent variables in our analyses. A latent variable is not directly measured or observed; its existence is inferred from other measures. In psychology the vast majority of our variables are latent – concepts such as emotions, moods, intelligence, aptitudes, beliefs or motivations are not amenable to direct measurement. Usually, we simply pretend that we have measured the variable that we are interested in, and do not distinguish between the variable that we measure and the variable in which we are interested: Self-esteem scale scores suddenly become 'self-esteem' itself. We sum the scores on our inventory, or we average the scores of two different observers, and we then treat this as if it were the variable we want. Rather than take this approach, structural equation models explicitly recognise that what we measure is not what we are directly interested in – rather we measure indicators of the latent variables.

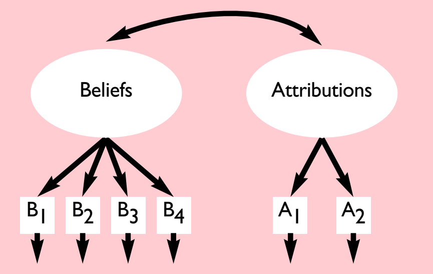
Thinking in terms of latent variables becomes useful in confirmatory factor analyses (CFA). In Figure 2 there are two latent variables – beliefs and attributions. The beliefs latent variable is measured using four observed variables. The attributions latent variable is represented by two observed variables. Each measured variable, or item, has two sources of variance – the latent variable, and measurement error. This is just the same as in a regression analysis. Now the latent variable represents only 'true' variance – the latent variable can be thought of as the variance that is shared among all of the measures. The latent variable represents the part of each item that we are interested in (actual beliefs), and removes the part we are not (errors, guessing, misunderstanding the question).

The correlation between the latent variable beliefs and latent variable attributions can be thought of as a correlation of the 'pure' measure, with the measurement error filtered out. In addition, by using a structural equation model we have tested the notion that the relationships amongst the measures of belief are explainable with a latent variable called 'beliefs'. This may or may not be the case, but SEM allows us to test this. A practical introduction to latent variable analysis is provided in Loehlin (1998).

Model fit

Structural equation models provide two types of inferential statistic. The first tests differences and relationships between different components of the model, in the form of standard errors (and so *p* values) around estimates of parameters in the model. The second type of inferential test is what sets SEMs apart from other forms of statistical tests. The statistic representing the overall fit of the model asks whether the model could have generated the data. Model fit is tested using a wide range of test statistics and is an area of considerable controversy (see Hu & Bentler, 1999, for a review and discussion of appropriate methods for testing fit). The main test of difference between model and data is the χ^2 (chi-squared) test.

FIGURE 2 A confirmatory factor analysis



Thinking about model fit requires us to think about inferential statistical tests rather differently from the way we usually think about them, for two reasons. The first reason is the source of the model. In conventional statistical analysis we start with our data, and we analyse the data to see what relationships there are in those data. We begin with data, and we move to the model. In SEM we begin with our model, and apply it to our data. Rather than asking 'What model will emerge from these data?' we ask 'Could these data have arisen from this model?' We are almost doing our statistical analysis backwards, beginning from where we want to be (the model) and seeing if we could have got there from where we were (the data).

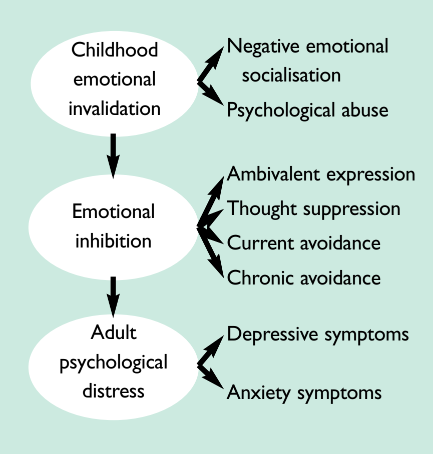
The second reason is the meaning of the fit. Our theories in psychology usually propose that there will be a difference between two groups or measures, or a relationship between two measures, and so we look for (and often hope for) statistical significance. SEM reverses this relationship – a statistically significant difference represents a difference between the model and the data, and SEM researchers usually hope for a statistically non-significant result.

Putting them together

The flexibility of SEM, the ability to use latent variables, and the power to test the fit of a given model result in a powerful technique for modelling psychological theory. Let's look at an example.

Krause *et al.* (2003) wanted to examine the relationship between childhood emotional invalidation (abuse and punishment) and adult psychological distress. They tested whether emotional inhibition could mediate the relationship between childhood emotional invalidation and adult distress. Each of these variables was considered to be a latent variable, which was measured using a number of observed variables, as shown in Figure 3. They found that the model was a good fit – there was no need to have a direct path from childhood emotional invalidation to adult psychological distress in order to account for the data, suggesting that childhood abuse does not directly influence current psychological distress: rather it does it by affecting emotional inhibition. (In these cases, the simpler model is considered the better model. In structural equation models, as in so much in life, small is beautiful.) Kraus *et al.* suggest that this may have implications for interventions to assist people who have suffered from emotional abuse – specifically the treatments should involve assisting with health expression of (particularly negative) emotion. It would be very difficult to provide this level of analysis using conventional techniques. In these cases the simpler model is considered the better model. In structural equation models, as in so much in life, small is beautiful.

FIGURE 3 Model of Krause *et al.* (2003)



Other recent examples of the use of SEM include Torkzadeh *et al*. (2003), who carried out a confirmatory factor analysis of computer self efficacy; Finkel *et al*. (2003), who used it to look at rates of decline in cognitive ability in old age, and to compare the rates of males and females; and Kupek (2002), who looked at memory errors and bias in the ability of people to recall the number of previous sexual partners.

Now we have convinced you...

We hope that having read this far you are now convinced that you should immediately go and find out more about SEM. A good overview of the approach is provided by Fife-Schaw (2000), and Ullman (2001) gives a practical introduction. But before you rush off to the library, we have to come clean about some of the drawbacks of the SEM approach.

The first problem that the budding structural equation modeller encounters is sample size. The sample sizes required for these analyses are large at best, and frequently very large. The factors that determine the sample sizes required can be complex, leading to difficulties in making appropriate statements about sample sizes (see Muthén & Muthén, 2002, for an approach to solving this problem). However, we would never be happy with a sample size of less than 100, but are considerably happier with a sample size in the hundreds.

Using SEM can add a sheen of respectability to a poorly executed research project, or a poorly thought-out analysis (Cliff, 1983; Cohen *et al*., 1990). Using SEM does not excuse the researcher from thinking about their research and analysis, and the theory behind their research. Indeed, it should encourage the researcher to think harder about their data and their analysis. SEM is not an exploratory technique, and cannot be (easily) used at the start of research into any field. The researcher must have a good idea of what they are looking for, at least in terms of appropriate variables, before they go ahead and collect and analyse data.

SEM can also suffer in the key scientific goal of replicability. Elsewhere in this issue Andy Field discusses the usefulness of meta-analysis in assimilating research findings. Models that are estimated using SEM are frequently difficult, or impossible, to assimilate. Different researchers may have used very different models, with different variables, and different sets of paths.

We conclude with a warning from David Rogosa, from the chorus to 'The Ballad of the Casual Modeler' (*www.stanford.edu/class/ed260/ballad.mp3*):

Now my model is busted
I can't make it fit
I drew in more arrows
But it still don't mean shit

■ *Jeremy Miles is in the Department of Health Sciences, University of York. E-mail: jnvm1@york.ac.uk.*
■ *Mark Shevlin is in the School of Psychology, University of Ulster. E-mail: m.shevlin@ulster.ac.uk.*

References

Barrett, P. (2002, 2 November). Re: Part 2: Exact fit vs. close fit testing [Item 029879]. Message posted to SEMNET discussion group, archived at bama.ua.edu/archives/semnet.html

Cliff, N. (1983). Some cautions regarding the application of causal modelling methods. *Multivariate Behavioral Research, 18*, 115–126.

Cohen, P., Cohen, J., Teresi, J., Marchi, M. & Velez, C.N. (1990). Problems in the measurement of latent variables in structural equation causal models. *Applied Psychological Measurement, 14*, 183–192.

Fekadu, Z. & Kraft, P. (2001). Predicting intended contraception in a sample of Ethiopian female adolescents: The validity of the theory of planned behavior. *Psychology and Health, 16*, 207–222.

Fife-Schaw, C. (2000). Introduction to structural equation modelling. In G.M. Breakwell, S. Hammond & C. Fife-Schaw (Eds.) *Research methods in psychology* (2nd edn, pp.397–413). London: Sage.

Finkel, D., Reynolds, C.A., McArdle, J.J., Gatz, M. & Pedersen, N.L. (2003). Latent growth curve analyses of accelerating decline in cognitive abilities in late adulthood. *Developmental Psychology, 39*, 535–550.

Heath, Y. & Gifford, R. (2002). Extending the theory of planned behavior: Predicting the use of public transportation. *Journal of Applied Social psychology, 32*, 2154–2189.

Hu, L. & Bentler, P.M. (1999). Cutoff criterion for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Krause, E.D., Mendelson, T. & Lynch, T.R. (2003). Childhood emotional invalidation and adult psychological distress: The mediating role of emotional inhibition. *Child Abuse and Neglect, 27*, 199–213.

Kupek, E. (2002). Bias and heteroscedastic memory error in self-reported health behavior: An investigation using covariance structure analysis. *BMC Medical Research Methodology, 2*, 14.

Loehlin, J. (1998). *Latent variable models: An introduction to factor, path and structural analysis* (3rd edn). Hillsdale, NJ: Lawrence Erlbaum.

Muthén, L.K. & Muthén, B.O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 4*, 599–620.

Torkzadeh, G., Koufteros, X. & Plughoeft, K. (2003). Confirmatory factor analysis of computer self-efficacy. *Structural Equation Modelling, 10*, 263–275.

Ullman, J.B. (2001) Structural equation modeling. In B.G. Tabachnick & L.S. Fidell (Eds.) *Using multivariate statistics* (4th edn, pp.709–811). New York: Allyn and Bacon.

Can meta-analysis be trusted?

ANDY P. FIELD *investigates the advantages and disadvantages of this method of analysing large amounts of data from diverse sources.*

UNTIL around 25 years ago the only way to assimilate and evaluate research evidence was through discursive literature reviews, in which someone with an interest in a given research topic would accumulate and subjectively evaluate the importance of research findings in that area. These reviews, although informative, are highly reliant on the discretion of the author who, with the best will in the world, could be unaware of important findings or could give particular importance to studies that others might believe to be relatively less important (see Wolf, 1986).

The failure of literature reviews to provide objective ways to assimilate scientific evidence led scientists to look for a statistical solution. The groundbreaking work of Glass (1976) and Rosenthal and Rubin (1978) paved the way for what we now know as meta-analysis: a statistical technique by which findings from independent studies can be assimilated.

The past 20 years

Meta-analysis is generally seen as an important step towards objectifying literature reviews; in fact, the extent to which meta-analysis is now regarded as an accurate and objective way to assimilate research findings is demonstrated by the proliferation of meta-analytic reviews in the major review journals in psychology (e.g. *Psychological Bulletin*). Figure 1 shows the number of published articles using or discussing meta-analysis that have appeared in peer-reviewed science journals over 20 years from 1981.

As you might expect, during the early 1980s when the technique was being honed by the likes of Hedges and Olkin (1985) and Hunter *et al*. (1982) the discussion of meta-analysis was, to say the least, sparse. Given a few years for the scientific community to absorb these seminal works (and that of Rosenthal, 1991), the use and discussion of meta-analysis suddenly rocketed from under 100 in the late 1980s to the several hundreds by the early to mid-1990s and to over a thousand by the turn of the century. A substantial proportion of these papers appears in social science journals.

Despite the obvious faith that psychologists and other scientists and practitioners have placed in meta-analysis, there is a growing body of evidence to suggest that it is often used incorrectly, and may not be the answer to our literature review prayers after all. This article describes the principles of meta-analysis before reviewing some of the sources of error that might make us doubt whether meta-analysis can be trusted.

Basic principles

As scientists, we measure effects in samples to allow us to estimate the true size of the effect in a population to which we don't have direct access (Field & Hole, 2003). Imagine I were interested in knowing the effect of sleep deprivation on people's ability to concentrate (and write articles about meta-analysis!). There is a true effect that sleep deprivation has, but I don't have access to that true effect because it's unlikely that I can sleep-deprive an entire population of people. Instead I use a small sample taken from that population and estimate the true effect that sleep deprivation has, based on the effect in my sample.

Now, the chances are that lots of other scientists will also be interested in the effects of sleep deprivation on concentration, and they too will have used samples to estimate the size of the effect that sleep deprivation has. The idea behind meta-analysis is simple: if we take all of these individual studies, quantify the observed effect in a standard way and then combine them, we can get a much more accurate idea of the true effect in which we are interested.

Effects can be quantified by expressing them as effect sizes (see David Clark-Carter's article in this issue): Cohen (1988) suggested a measure called *d*, but we can use the Pearson correlation coefficient *r*, odds ratios or risk rates. Typically, the choice of measure depends on the conventions of the research discipline and is not based on statistical reasoning. For example, the correlation coefficient is typically chosen to represent the size of a relationship, and Cohen's *d* is used to quantify the degree of difference between group means; however, the correlation coefficient, Pearson's *r*, can be used to quantify differences between means (see Field & Hole, 2003; Rosenthal, 1991). All effect size estimates represent a standardised form of the size of the observed effect, and most can be easily transformed into a different metric and back again (see Rosenthal, 1991); however, there are often statistical reasons for preferring one metric to another.

The first step in meta-analysis is, therefore, to express the effect in each study in a uniform way. So, if we decide to use *r* as our effect-size measure, we would need to look at each study and use the data to calculate the value of *r*. The mean of these effect sizes can then be calculated. In addition, although this isn't always the primary concern of meta-analysis, the probability of obtaining that mean can also be computed (see Field, 2000). In short, we can see whether the average effect size is significant, which, to use my sleep deprivation example, would tell us the average size and statistical significance of the relationship between sleep deprivation and concentration across several studies (from which the size of the effect of sleep deprivation on concentration in the population can be inferred).

Meta-analysis can also be used to find out the variability between effect sizes across studies (called tests of the homogeneity of effect sizes) and to explain this variability in terms of moderator variables. For example, we might find that in a study that tested concentration using a visual task, the effect of sleep deprivation was larger than in a different study that tested concentration using an audio task. If we had several studies using visual tasks and several using audio ones, then we could test whether there was significant variability across effect sizes, and also test whether this variability is caused by the modality of the task (audio vs. visual) – see Field (2003).

Publication bias and the 'file drawer' problem

A major threat to the validity of meta-analysis is that significant findings are more likely to be published than non-significant findings. This is both because researchers may lose interest in non-significant findings and not submit them (Dickersin *et al*., 1992) and because reviewers may scrutinise non-significant findings more closely and reject manuscripts containing them (Hedges, 1984).

Rosenthal (1979) calls this the 'file drawer' problem – non-significant research is more likely to end up in the researcher's file drawer than in a journal! The extent of this problem should not be underestimated: Sterling (1959) reported that 97 per cent of articles in psychology journals reported significant results, Greenwald (1975) has estimated that significant findings were eight times more likely to be submitted than non-significant ones, and unpublished research can have effect sizes half the value of comparable published research (Shadish, 1992).

The effect of this bias is that meta-analytic reviews are likely to overestimate mean effect sizes (and their significance) because they might not include unpublished studies, in which effect sizes would have been small.

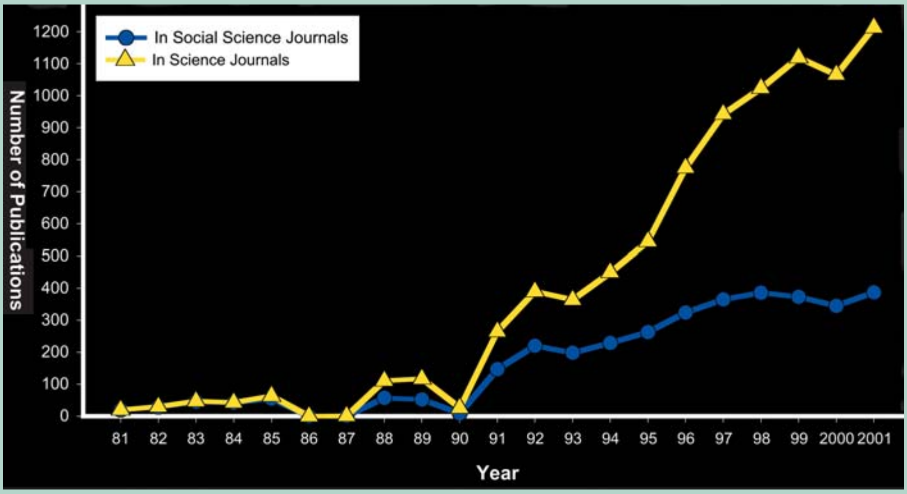
Artefacts

Many different artefacts may also introduce error into a meta-analysis. Although some artefacts may be unique to certain research disciplines, those that stem from the measurement of variables and the general quality of the research apply in all situations. The accuracy of the effect size of a variable will depend largely on how accurately that variable was measured and on how likely it is that the error in the measurement of variables will vary across studies. For example, one study might have used a very reliable questionnaire, whereas another uses a less reliable one. In addition, correlational research studies can vary in the range of scores elicited from participants (range variation); these differences in the range of scores elicited will affect the resulting effect sizes.

More generally, we can say that effect sizes are influenced by the quality of the research. In its simplest form, meta-analysis doesn't take account of the measurement reliability, range differences, or the general quality of research. Although Hunter and Schmidt (1990) have suggested statistical techniques for correcting for measurement error and range variation, many researchers either do not apply these corrections or apply them incorrectly (Schmidt & Hunter, 1996).

Given that artefacts contribute to differences in effect sizes, an alternative approach is to look for significant variability between effect sizes in a meta-analysis (and we've seen earlier that this is possible). If significant variability is found, then potential moderator variables can be sought to explain this variability. One possible moderator might be the general quality of the research, and Wolf (1986) suggests that the quality of research can be used as a moderator variable that tests whether the effect size is significantly different in 'well-conducted' and 'badly-conducted' studies. However, this introduces subjective opinion about what is well-conducted research and so perhaps makes the analysis no more objective than the discursive evaluations that meta-analysis seeks to objectify! To make matters worse, although many researchers do test for variability between effect sizes, they rarely act upon the results of these tests (Hunter & Schmidt, 2000).

FIGURE 1 Number of articles about or using meta-analysis published in science journals 1981–2001 (including those appearing specifically in social science journals). Source: *wos.mimas.ac.uk*.



Misapplications

Another implication of variability between effect sizes relates to how the meta-analysis is conceptualised and calculated. There are two ways to conceptualise meta-analysis: fixed-effects and random-effects models (there is actually a mixed model too, but for simplicity I'll ignore it). These models differ not only in the theoretical assumptions that underlie them, but also in how the mean effect size and its significance is computed. The fixed-effects model assumes that all studies in a meta-analysis come from a population in which the average effect size is fixed (Hunter & Schmidt, 2000). The alternative assumption is that the average effect size in the population varies randomly from study to study, so population effect sizes can be thought of as being sampled from a 'superpopulation' (Hedges, 1992).

Statistically speaking, the main difference between fixed- and random-effects models is in the amount of error.

In fixed-effects models there is error introduced because of sampling studies from a population of studies. This error exists in random-effects models but in addition there is error created by sampling the populations from a superpopulation. So, calculating the error of the mean effect size in random-effects models involves estimating two error terms, whereas in fixed-effects models there is only one error term.

Another reason why meta-analysis might not be trusted is that the wrong model may be used. There is considerable theoretical (Field, 2003; Hunter & Schmidt, 2000; National Research Council, 1992) and empirical (Barrick & Mount, 1991) evidence that real-world data are likely to reflect the random-effects conceptualisation (that is, studies come from populations in which the average effect size varies). However, Hedges and Vevea (1998) suggested that the choice of model depends not on the assumptions about the true state of the world, but on the type of inferences that the researcher wishes to make: with fixed-effects models inferences can be drawn only about the studies included in the meta-analysis whereas random-effects models allow inferences that generalise beyond the studies included in the meta-analysis. Psychologists typically wish to generalise beyond the studies included in the meta-analysis, so random-effects models are more appropriate (see Field, 2003; Hunter & Schmidt, 2000).

Despite good evidence that real-world data support a random-effects conceptualisation, the relative simplicity of fixed-effects models has meant that psychologists routinely apply them – even though the random-effects model is more often appropriate. In fact, even when tests reveal significant variability between effect sizes (suggesting a random-effects model should be used), psychologists do not act upon these tests and apply fixed-effects models regardless. Hunter and Schmidt (2000) found 21 recent examples of meta-analytic studies using fixed-effects models in *Psychological Bulletin* (the highest-impact review journal for psychology) compared with none using random-effects models.

What are the consequences of misapplying fixed-effects models to random-effects data? Well, in short, it inflates the estimate of the mean effect size and its significance: normally, using a standard criterion for significance ($p < .05$), we would expect to find a significant average effect size, when there is no effect in the population, in around 5 per cent of cases (the Type I error rate). Hunter and Schmidt (2000) predict that this error rate will increase to between 11 per cent and 28 per cent of cases. However, using data simulations I have shown (Field, 2003) that in fact the error rates increase to anywhere between 43 per cent and 80 per cent in certain circumstances. To put this into perspective, of the 21 meta-analyses reported by Hunter and Schmidt (2000), between 9 and 17 of them are likely to have reported significant average effect sizes when, in reality, no significant effect existed within the population.

Three drug tests in three countries, three effects – How would a meta-analysis distort these results?

...and more seriously

A more fundamental issue is that, given that we assume real-world data follow a random-effects model and effect sizes vary across studies, then what is the value in seeking an average effect size?

For example, imagine we tested the efficacy of a powder ('Stat-Whizz') that could magically make you good at statistics. A trial in the US found an effect size of .45, a replication in Belgium found an effect size of 0, and a further replication in the UK yielded an effect size of –.45. If we assume that these studies had equal sample sizes and so were equally weighted in the meta-analysis, then the resulting average effect size would be 0 – there would be a non-significant effect.

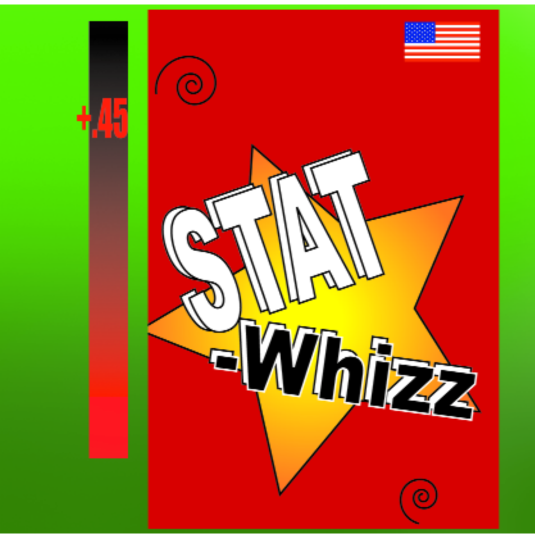
Readers of such a meta-analysis might conclude, therefore, that Stat-Whizz was an ineffective drug. Of course, this conclusion is wrong: the drug worked in the US, didn't work in Belgium and had a negative effect in the UK. The issue of interest is not so much the overall effect of the drug, but at what levels the drug works: the fact that the drug doesn't work on the English is of little interest to all of the Americans for whom the drug is effective! A retort to this is that such variability would be picked up by tests of the homogeneity of effect sizes. However, as mentioned previously, researchers frequently fail to act upon such tests (Hunter & Schmidt, 2000).

One implication of these observations is that moderator analysis, in which we look for possible variables that explain the variation between effect sizes, may be more useful than looking at average effect sizes.

Methods of meta-analysis

A final possible source of error in meta-analysis could be problems inherent in the method used. Three methods of meta-analysis have been popular: the methods devised by Hedges and colleagues (Hedges 1992; Hedges & Olkin, 1985; Hedges & Vevea, 1998), and the methods of Rosenthal and Rubin (1978) and Hunter and Schmidt (1990). Hedges and colleagues have developed both fixed- and random-effects models for combining effect sizes, Rosenthal and Rubin have developed only a fixed-effects model, whereas Hunter and Schmidt label their method a random-effects model. The computations of these various methods differ, and the technical details of these differences are well documented elsewhere and are beyond the scope of this review (see Field, 2001).

Several recent studies have compared these methods. Johnson *et al*. (1995) compared the Hedges–Olkin (fixed-effect), Rosenthal–Rubin and Hunter–Schmidt meta-analytic methods by manipulating a single data set. They concluded that the



significance of the mean effect size differed substantially across the methods: the Hunter and Schmidt method reached more conservative estimates of significance than the other two methods, so should be used cautiously. Schmidt and Hunter (1999) subsequently claimed that Johnson *et al*. incorrectly applied their method and showed that, theoretically, when the method was correctly applied, their method was comparable to that of Hedges. I highlighted some other concerns with the methods of Johnson *et al*. and rectified these concerns in a series of simulations that compared the methods across a variety of situations (Field, 2001).

When comparing random-effects methods, the Hunter–Schmidt method yielded the most accurate estimates of population effect size across a variety of situations. However, neither method controlled the Type I error rate when 15 or fewer studies were included in the meta-analysis.

The method described by Hedges and Vevea (1998) controlled the Type I error rate better than the Hunter–Schmidt method when 20 or more studies were included. In a more recent set of simulations (Field, 2002) I demonstrated that across a far-ranging set of situations both methods produce biased estimates of the population effect size: however, the biases in the Hunter–Schmidt method are not as large as in the Hedges method. Hedges' method did tend to keep tighter control of the Type I error rate, but with 80 or more studies in the meta-analysis, there was little to separate the two methods. With fewer studies in the meta-analysis (20–40), Hedges' method controlled the Type I error rate considerably better than Hunter and Schmidt's method. As a general rule, neither method was accurate when fewer than 20 studies were in the meta-analysis.

So, can it be trusted?

To sum up, meta-analysis has come to be seen as the saviour of the literature review, but perhaps unjustly. As with all statistical procedures, the results are only as good as the data available and the person performing the test: if fixed-effects models continue to be routinely applied to psychological data, then we risk finding inflated effects. In terms of which method to apply, if the primary interest is in estimating the effect in the population, then what matters is whether it's better, in the context of the question you're trying to address, to underestimate (Hunter–Schmidt) or overestimate (Hedges) the effect size in the population. If the significance of this estimate is important, and there are 80 or more studies in the meta-analysis then either method will be fairly reliable, but with 20–40 studies Hedges' method is preferable, and significance tests should not be conducted at all with fewer than 20 studies in the meta-analysis.

Of course, there is more to meta-analysis than this simple summary suggests. I've hinted at the fact that moderator variables may often be more interesting than the average effect size. Also, researchers have to give greater consideration to controlling for other sources of error – small statistical differences between the methods discussed here may be relatively unimportant compared with biases from other sources, such as using unreliable measures.

■ *Dr Andy P. Field is in the Department of Psychology, University of Sussex. Tel: 01273 877150; e-mail: andyf@cogs.susx.ac.uk.*

References

Barrick, M.R. & Mount, M.K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1–26.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd edn). Hillsdale, NJ: Lawrence Erlbaum.

Dickersin, K., Min, Y-I. & Meinert, C.L. (1992). Factors influencing publication of research results: Follow-up of applications submitted to two institutional review boards. *Journal of the American Medical Association, 267*, 374–378.

Field, A.P. (2000). *Discovering statistics using SPSS for Windows: Advanced techniques for the beginner*. London: Sage.

Field, A.P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*, 161–180.

Field, A.P. (2002). *Is the meta-analysis of correlation coefficients accurate when population effect sizes vary?* Manuscript submitted for publication.

Field, A.P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics, 2*, 77–96.

Field, A.P. & Hole, G. (2003). *How to design and report experiments*. London: Sage.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 51*, 3–8.

Greenwald, A.G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.

Hedges, L.V. (1984). Estimation of effect size under non-random sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*, 61–85.

Hedges, L.V. (1992). Meta-analysis. *Journal of Educational Statistics, 17*, 279–296.

Hedges, L.V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L.V. & Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486–504.

Hunter, J.E. & Schmidt, F.L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.

Hunter, J.E. & Schmidt, F.L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative knowledge in psychology. *International Journal of Selection and Assessment, 8*, 275–292.

Hunter, J.E., Schmidt, F.L. & Jackson, G.B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Johnson, B.T., Mullen, B. & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology, 80*, 94–106.

National Research Council (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.

Rosenthal, R. (1979). The 'file drawer' problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. edn). Newbury Park, CA: Sage.

Rosenthal, R. & Rubin, D. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavior and Brain Sciences, 3*, 377–415.

Schmidt, F.L. & Hunter, J.E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*, 199–223.

Schmidt, F.L. & Hunter, J.E. (1999). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, and Salas (1995). *Journal of Applied Psychology, 84*, 144–148.

Shadish, W.R. (1992). Do family and marital psychotherapies change what people do? A metaanalysis of behavioural outcomes. In T.D. Cook, H. Cooper, D.S. Cordray, H. Hartmann, L.V. Hedges, R.J. Light *et al*. (Eds.) *Meta-analysis for explanation: A casebook* (pp.129–208). New York: Sage.

Sterling, T.C. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association, 54*, 30–34.

Wolf, F.M. (1986). *Meta-analysis: Quantitative methods for research synthesis*. Newbury Park, CA: Sage.

How to… Produce a bad results section

You worked hard for your data – why share them all? **DANIEL B. WRIGHT** *and* **SIÂN WILLIAMS** *come to the rescue.*

COMPLICATED equations, confusing figures, arcane technical expressions; all commonly found in psychology results sections. In this article we'll show you how to achieve these lofty heights of mind-numbingly boring techno-babble. Inappropriate use of statistical procedures, bad graphs, poor writing style... we'll cover the lot. Your findings will be so obscure that even you won't understand them.

Our approach follows Howard Wainer (1984), who described how to make graphs as uninformative as possible. His approach was to 'concentrate on methods of data display that leave the viewers as uninformed as they were before seeing the display or, worse, those that induce confusion' (p.137). But perhaps Wainer didn't go far enough: we show how entire results sections can be made to 'induce confusion'. Many authors of results sections published in the most respected journals already recognise the value of obscurity. Indeed, the American Psychological Association created a task force to discover why so many follow our approach (Wilkinson *et al*., 1999).

Statistical tests: Failing the four Rs

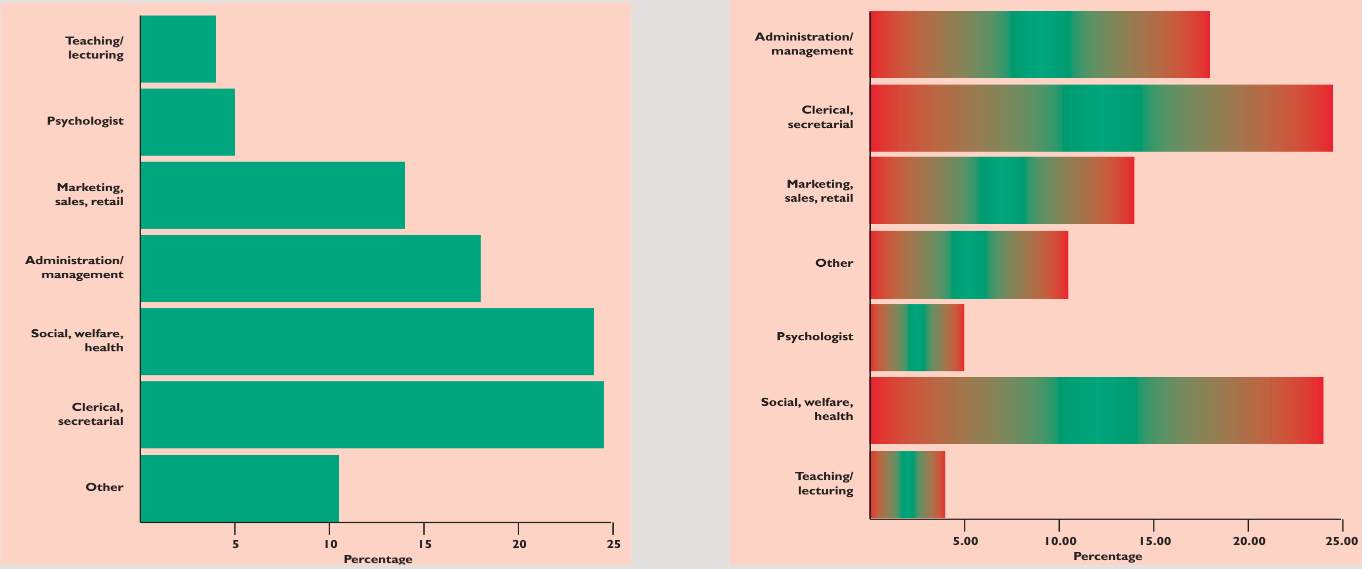
There are three basic ways to miscommunicate findings: numerical, graphical and verbal. The first, numerical, relates to the statistical tests that people conduct.

To produce a results section that is completely misleading, the author could conduct statistical tests that are clearly wrong. Easily achieved when you consider that when questioned, many researchers do not even understand concepts fundamental to much of the statistics that they use, like what *p* means or what a confidence interval is (Oakes, 1986). But, unfortunately, reviewers have a tendency to notice when wrong tests are used. A subtler tactic that can often get past reviewers is only conducting and reporting the final hypothesis-testing statistics, and not exploring the data. Failing to explore the data adequately can mean that interesting facets of the data will not be discovered by the researcher and thus will be hidden from the reader.

Hoaglin *et al*. (1983) discuss the four Rs of understanding data:

- *Resistance* Some statistics are not 'resistant' – they are heavily influenced by a small fraction of the data. (This concept is closely related to a statistic being robust. Resistance is a characteristic of robust statistics.) The mean, ANOVAs, ordinary regressions, and so on, are not resistant, so you should use them (or you could always see Wilcox, 2001, for an introduction to some alternatives).
- *Residuals* This refers to how different

FIGURE 1 Destination of psychology graduates in the UK: same data, different presentation styles



points do not fit with the model. Much as Piaget showed how focusing on children's errors could shed light on cognitive development, it is necessary to examine the residuals to judge the worth of any model.

- *Re-expression* Should the raw data be rescaled to make the analyses and interpretation simpler? Often this is to make the scale of the data more appropriate for the theories being investigated or to make the data more resistant (usually more like the normal distribution). For example, reaction time data are often transformed so that the distribution is not as positively skewed.
- *Revelation* Your methods of analysis can often reveal interesting and unexpected aspects of the data and help inform theories. Following our advice should limit this possibility.

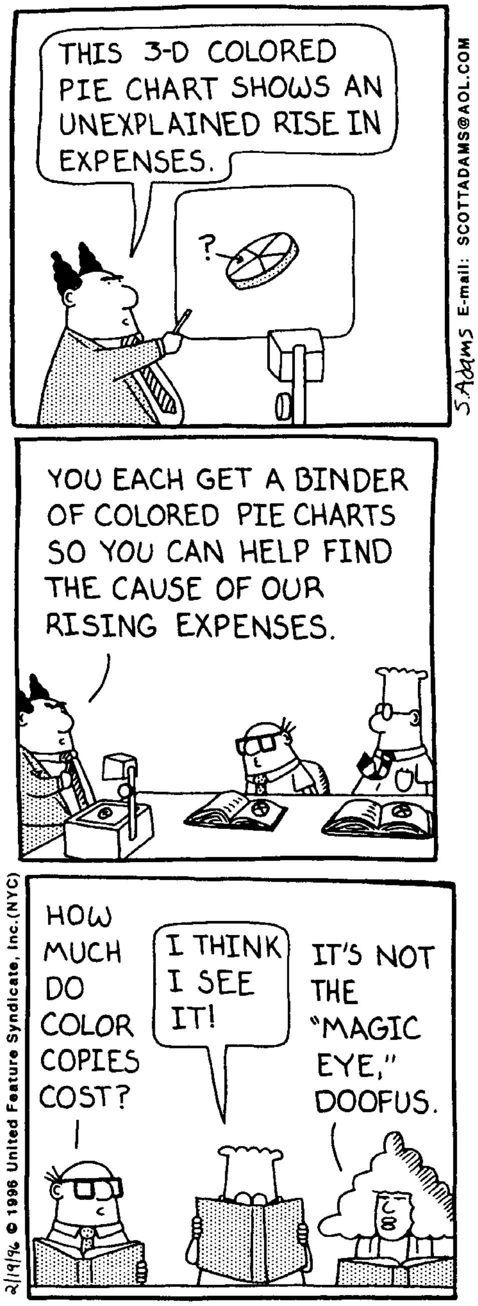
As our goal is to help people create misleading and uninformative results sections, we recommend ignoring the four Rs. Instead, find an introductory textbook that has a flowchart that presents simple questions like 'Are you interested in an association, or group differences?' and 'What is the level of measurement of the data?' and directs the reader to one particular test. This approach, used on its own and with little consideration of the questions, should leave you with just an *r*, *F* or χ^2 value as the only means to decide whether the statistical model being considered is appropriate. Don't bother with graphing and examining the descriptive statistics before performing any inferential statistics.

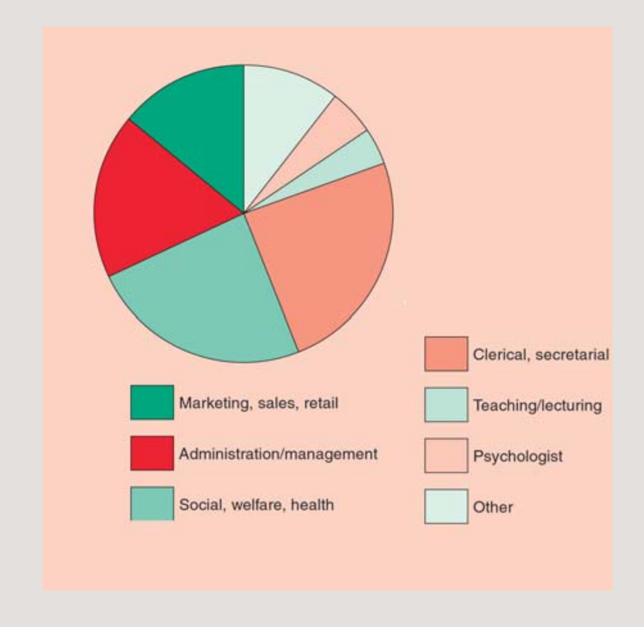
Making bad graphs

For decades the science of graphical display developed so that politicians could construct misleading graphs, with the assumption that the audience was not interested in the numbers unless they were made artistically appealing. Tufte (2001) and Wainer (1984) show some graphs, from highly respected sources, that hide the data from readers, display data inaccurately and present them in a cluttered and confused manner. Many of these methods have arisen because computers can add extra frills to graphs – what is called 'chartjunk'. While computers have made making graphs that clearly communicate the findings easier, they have also created a fertile environment for you to smother your data with technology: 'like weeds, many varieties of chartjunk flourish' (Tufte, 2001, p.107).

As an example, Figure 1 shows three graphs giving the destinations of psychology graduates in the UK in 1999 (data from the BPS document *Studying Psychology*, 2001). In the first graph the reader can see the frequency increasing from 'teaching/lecturing' to 'clerical, secretarial', but the reader won't know that with most graphics software you can tick lots of fancy options. In the middle graph we've alphabetised the items, added a really neat visual illusion to the bars and altered the axis labels. Looks pretty and makes the information more opaque. Many of the patterns that can be used to fill bars create visual illusions about the size, the shape and even the apparent motion of the bars. The third graph is a pie chart, which generally makes it more difficult for the reader to extract information (Hollands & Spence, 2001). Then the boss from Dilbert shows (or he'd say 'demonstrates') the pinnacle of bad graphs. According to the SYSTAT manual, these false 3–D pie charts 'incorporate nearly every visual illusion discussed in this chapter' (Wilkinson, 2000, p.13).

Being creative with colour can further demonstrate how technology can triumph over communication. Consider using red and green, to make the graph particularly bad for the approximately 10 per cent of males who have difficulty distinguishing these colours.


THIS 3-D COLORED PIE CHART SHOWS AN UNEXPLAINED RISE IN EXPENSES.
?
S.Adams E-mail: SCOTTADAMS@AOL.COM
YOU EACH GET A BINDER OF COLORED PIE CHARTS SO YOU CAN HELP FIND THE CAUSE OF OUR RISING EXPENSES.
HOW MUCH DO COLOR COPIES COST?
I THINK I SEE IT!
IT'S NOT THE "MAGIC EYE," DOOFUS.
2/19/96 © 1996 United Feature Syndicate, Inc.(NYC)



Miscommunicating with the written word

Even if you have written a brilliant literature review, have meticulously and clearly described your study, and have written a fluent and informative discussion, there are still several ways to make your paper 'induce confusion'.

Assume all rules of writing style are not applicable for results sections. Many readers expect results sections to be confusing. This is because the standards of good writing do not apply to results sections. In their book *The Elements of*

Style Strunk and White (1979) argue the case for 'cleanliness, accuracy, and brevity' (p.xiii) in the use of language. Messiness, inaccuracy and extravagance will ensure that people avoid even reading the results section.

Sometimes it is worth doing more than just inducing confusion. Sometimes it is worth producing annoyance. Most scientists are driven by a sense of curiosity (Grigorenko, 2000). We question things. You can feed this curiosity by reporting only some of the information. For instance, omitting the degrees of freedom for an analysis may leave your audience curious about your sample size. It is also possible, and very easy, to neglect to describe any transformations you have made to the data, or how you dealt with any outliers. However, our favourite method is conducting a statistical test, say a *t* test, and not reporting the means. Imagine telling someone in a pub about a study and instead of saying 'the group who were given the drug answered, on average, 10 per cent more questions correctly, and this much of a difference would be unlikely if the drug had no effect', you said '$t(18) = 2.30$' and nothing else. The American Psychological Associations task force on statistical inference (Wilkinson *et al*., 1999) stressed the importance of reporting descriptive statistics. If you wish to produce a bad results section, ignore everything in their report.

Try to make the statistical techniques sound as complicated as possible. Some statistical techniques are very complex. Assume that the reader has a PhD in statistics and knows every statistical technique. Use lots of jargon, particularly if you are using some esoteric technique that you have just learnt. Explain every minute mathematical aspect of the technique. This is easily done by paraphrasing statistics books and manuals – you don't need to understand them yourself. Use a thesaurus to slightly change the meanings of words so that it is not copying word-for-word from the manual. Given that some words have precise technical meanings, this should also confuse readers who felt that they understood the techniques. For example, the words *components* and *factors* have different meanings, and result from different statistical procedures, but are sometimes interchanged.

Use the word *significant* as if it meant the effect was large and important. Some words have a different meanings in English and Statisticalese. Usually the words have the same basic meaning, but have a more precise definition in scientific jargon than in English. But sometimes the meanings are very different. *Significant* in Statisticalese means that assuming the null hypothesis is correct, data as extreme as observed should occur less than 5 per cent of the time. The word *significant* means something very different in English: important. One way to confuse readers is to assume being statistically significant means that the effect is significant in the English sense of the word.

Be careless about using causal language when describing correlational studies. Both causal and associative hypotheses are important in psychology. However, they are different with respect to the theories that are being investigated, and they require different research designs. We recommend casually using words like *cause* and *influence* when conducting studies where there has been no manipulation.

But if you insist on doing it properly…

Irony aside, many of the techniques that we have shown can be easily avoided. There were several themes running through this article.

For analysis, use exploratory techniques to understand the data before you leap into statistical tests. Become friends with your data (Wright, 2002, 2003). Don't just check if a result is statistically significant: look at the size of the effect, ask if it is robust, check to make sure that it is consistent with your graphs, and ask if your finding makes sense.

For graphs, showing technical wizardry and making them 'pretty' can be at odds with the aim of clearly and accurately communicating results. See Wainer and Velleman (2001), and Tufte's marvellous trilogy *The Visual Display of Quantitative Information* (2001, first edition published 1983), *Envisioning Information* (1990), which is about picturing nouns, and *Visual Explanations* (1997), which is about picturing verbs.

For writing styles, think about your audience: second-year undergraduates should be able to understand what you write. Strunk and White (1979) and Sternberg (2000) highlight numerous ways in which writing styles can be used to improve the presentation of research findings.

The art of conducting and communicating statistics is difficult. Abelson (1995) describes how people should consider what they are trying to persuade the reader about. He gives five MAGIC criteria (Magnitude, Articulation, Generality, Interestingness and Credibility) that you should bear in mind whenever you are reporting a statistical result. You should be excited by your results and convey this to the reader. If you are bored by your results, your readers will be too. If scientific papers were murder mysteries, the results section would reveal the killer.

■ *Daniel B. Wright is in the Psychology Department at the University of Sussex. E-mail: danw@sussex.ac.uk.*

■ *Siân Williams is at the Royal College of Paediatrics and Child Health. E-mail: sian.williams@rcpch.ac.uk.*

References

Abelson, R.P. (1995). *Statistics as principled argument*. Hillsdale, NJ: Lawrence Erlbaum.

Grigorenko, E.L. (2000). Doing data analyses and writing up their results: Selected tricks and artifices. In R.J. Sternberg (Ed.) *Guide to publishing in psychology journals* (pp. 8–120). Cambridge: Cambridge University Press.

Hoaglin, D.C., Mosteller, F. & Tukey, J.W. (1983). *Understanding robust exploratory data analysis*. New York: Wiley.

Hollands, J.G. & Spence, I. (2001). The discrimination of graphical elements. *Applied Cognitive Psychology, 15*, 413–431.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. Chichester: Wiley.

Sternberg, R.J. (Ed.) (2000). *Guide to publishing in psychology journals*. Cambridge: Cambridge University Press.

Strunk, W. Jr & White, E.B. (1979). *The elements of style* (4th edn). Needham Heights, MA: Allyn & Bacon.

Tufte, E.R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Tufte, E.R. (1997). *Visual explanations: Images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

Tufte, E.R. (2001). *The visual display of quantitative information* (2nd edn). Cheshire, CT: Graphics Press.

Wainer, H. (1984). How to display data badly. *American Statistician, 38*, 137–147.

Wainer, H. & Velleman, P.F. (2001). Statistical graphics: Mapping the pathways of science. *Annual Review of Psychology, 52*, 305–335.

Wilcox, R.R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer-Verlag.

Wilkinson, L. (2000). Cognitive science and graphic design. In SPSS Inc., *SYSTAT 10: Graphics* (pp.1–18). Chicago, IL: SPSS Inc.

Wilkinson, L. and the Task Force on Statistical Inference, APA Board of Scientific Affairs (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604.

Wright, D.B. (2002). *First steps in statistics*. London: Sage.

Wright, D.B. (2003). Making friends with your data: Improving how statistical results are reported. *British Journal of Educational Psychology, 73*, 123–126.