

## SPECIAL ARTICLE

# *All that glisters...* How to assess the 'value' of a scientific paper

J. J. Pandit<sup>1</sup> and S. M. Yentis<sup>2</sup>

*1 Consultant Anaesthetist, Nuffield Department of Anaesthetics, John Radcliffe Hospital, Oxford OX3 9 DU, UK*

*2 Consultant Anaesthetist, Magill Department of Anaesthesia, Intensive Care and Pain Management, Chelsea and Westminster Hospital, 369 Fulham Road, London SW10 9NH, UK*

---

Correspondence to: J. J. Pandit

E-mail: [jaideep.pandit@physiol.ox.ac.uk](mailto:jaideep.pandit@physiol.ox.ac.uk)

Accepted: 26 December 2004

Doctors are expected to keep up to date with the literature, and to change their practice on the basis of what they have read. We believe that the skills needed for evaluating and interpreting scientific papers are often lacking. Even when they are taught, the emphasis is often focused on technical aspects of a paper without consideration of its *value*, a term we use here to encompass not only the quality of a paper's methods, but also its context and importance. We argue that *critical appraisal*, the term used to describe the assessment of a paper's methodological niceties [1], should never take place in isolation but must always occur in parallel with assessment of its value, since a paper may be methodologically sound but contribute little to a better understanding of the subject. Here, we discuss various aspects of scientific papers that contribute to or detract from their value, and suggest a stepwise approach for assessing them.

### Background: the scientific method

The purpose of a scientific publication, of which there are numerous types, is to communicate information. *Editorials* are summary or personal views, perhaps commenting on a specific paper. *Reviews* are longer, in-depth analyses of the literature; they are most persuasive when their analysis is objective but their conclusions (as with editorials) are often subjective and personal. Specific forms of supposedly objective reviews are *quantitative* or *systematic reviews*, or *meta-analyses* [2]. *Case reports* describe one (or more) specific clinical cases, either so unusual as to be of interest, or from which it is hoped some particular lesson can be learned. A *letter* (correspondence) is a comment on, or criticism of, another's published paper, or seeks answers to a question. A letter may also convey information or data

that do not constitute a full study but nonetheless are thought worthy of dissemination.

Finally, there are *experimental investigations*. Broadly, these attempt to answer a question by the use of the 'scientific method', which involves the following process [3, 4]. First, the researcher formulates a hypothesis. This hypothesis may represent current ideology (the current theory, or *paradigm*), or it may be an idea developed *de novo* (e.g. suggested by preliminary observations). The hypothesis leads to an *experimental prediction*: if a certain experiment is conducted, then this predicted result should obtain if the hypothesis is correct. A result consistent with the prediction supports the hypothesis, but does not 'prove' it (thus scientific proof is very different from mathematical proof). If, on the other hand, the experimental result is not as was predicted, then either the hypothesis is incorrect (so disproved) or the conduct of the experiment was flawed.

There are many types of experimental investigations: clinical studies or trials (the terms are often used interchangeably, although the latter is sometimes restricted to investigations of a treatment's efficacy); laboratory investigations; mathematical modelling of data; and some observational studies and audits. We focus here on experimental investigations since these are the papers that advance the knowledge base the most.

### Essential components of an experimental investigation

There are generally two aspects to excellence in an experimental study: first, the study's *conduct* (of which handling of the data is an inherent part), and second, its *presentation*. Important aspects of presentation are covered

by journals' instructions to authors and by specific publications [5, 6], and we will not discuss these in detail here.

### Conduct of the study

First, the study must consider a clear hypothesis that is stated unambiguously, ideally, illustrated by the results that would be expected if the hypothesis were true. The final results and conclusion of the study should relate to this prediction and must either refute or be consistent with the hypothesis.

Second, the study must have appropriate ethical approval (or conform to animal research guidelines). We will not consider this issue further, since it has been addressed recently [7].

Third, the technical conduct of the experiment must be sound. Particular attention should be given to the avoidance of surrogate measures, appropriate measurement tools, proper randomisation and blinding, appropriate use of control groups, and appropriate application and interpretation of statistics. We consider each of these below.

#### Surrogate measures

Surrogate measures or end-points have serious limitations. For example, a study of the control of minute ventilation may not actually involve measurement of this outcome at all, but instead may try to derive conclusions based on the measurement of another, related variable (say, arterial  $PCO_2$ ). The problem is that other factors may influence the related variable (e.g. ventilation is not the only factor influencing  $PCO_2$ ). Furthermore, measures that seem superficially related may not be; for example, although flecainide reduces cardiac arrhythmias, it increases mortality – a much more relevant end-point [8]. Surrogate measures are sometimes used in clinical studies because of the difficulty in measuring the desirable end-point, such as long-term survival. They might also give crude estimates of trends over time for certain variables [9], but they have very little (if any) place in studies that seek to question or overturn fundamental hypotheses in the underlying science.

#### Measurement tools

Measuring devices and assessment tools must be valid (measure what they are supposed to measure), accurate (measure the true quantity), and reliable (different users obtaining the same results) – aspects that often escape attention in manuscripts. This is not restricted to technical measurements; for example, assessment of 'maternal satisfaction' with the use of a simple visual analogue scale continues widely despite little evidence to support it [10]. When technology is used, coefficients of variation should

be given, but rarely are. Further, there ought to be some confidence as to how the technology works. For example, much of the growing literature on the bispectral index (BIS) as a monitor of 'awareness' raises concerns. Since it is not known precisely what is being measured by the BIS [11], it can never be known whether an unexpected result has arisen because the BIS is invalid or inaccurate, or because the hypothesis being tested is incorrect [12].

#### Randomisation and blinding

Randomisation and blinding are intended to minimise the influence of bias. The hope with randomisation is that all 'confounding factors' (both known and unknown) that might influence the outcome will be distributed equally amongst the groups. If any differences are found they can therefore be attributed to the sole factor – the treatment under study – that has not been 'shared out' in this way. However, even with proper randomisation, groups may be unequal: chance alone might result in one group's subjects being older, heavier, younger or just luckier than those in the other group. Even when groups appear equal, small inequalities might combine to influence the results. For example, in a study by Greif *et al.* [13] into the possible anti-infective effect of peri-operative oxygen therapy, patients randomly allocated to receive extra oxygen were by chance more likely to be fitter and less likely to be smokers, to have inflammatory bowel disease, and to undergo rectal surgery than those in the 'no oxygen' group. Could these factors have combined to contribute to the dramatic reduction in infection seen in the 'oxygen' group, such that a subsequent study obtained completely the opposite results [14]? In addition, the human urge to guess or manipulate treatment allocations, or otherwise interfere with proper randomisation in studies, is well reported [15].

Blinding is present when the person treating the patient, or making the assessments, does not know which patients receive which treatment. Some studies fail to take even the simplest steps to ensure blinding, while others go to extraordinary lengths (for an example of the latter, see Smith and Thwaites's [16] commendable study comparing intravenous with inhalational anaesthesia). Occasionally, blinding is impossible (e.g. when comparing two different laryngoscope blades or bougies [17]), but the results of a blinded study are always more persuasive. Indeed, studies with insufficient blinding tend to report greater treatment effects than those with proper blinding procedures [15].

#### Control groups

Control groups may be inappropriate because of poor randomisation or blinding. Even if these are sound, though, there may be other reasons why treatment effects may be masked or exaggerated by problems with control

groups: lack of consideration of other possible 'confounding factors'; use of historical, rather than contemporaneous, controls; comparison of a treatment against placebo instead of standard practice, or against an inappropriate treatment; or (at worst) lack of a control group at all.

#### Statistics

Much of the above might give the impression that the testing of hypotheses by close attention to the study's conduct will always give a clear-cut result. Unfortunately, this is not the case, because we can never achieve certainty; the best we can do is use statistical analysis to indicate the degree of uncertainty [18]. A full discussion of statistical methods is dealt with elsewhere [19], and here we consider only two related areas that commonly cause difficulties: significance and power.

*Significance.* If, in a study, one drug appears more effective than another, the traditional approach is to ask the question: 'What is the likelihood (or probability) that this result is a chance finding, and that these two drugs are in fact equivalent?' This approach (testing the equivalence, as opposed to testing the difference) is known as testing the null hypothesis. It is important to stress that this null hypothesis assessed by the statistical test may not always be exactly the same as the underlying scientific hypothesis being examined by the study as a whole: the result of the former will help interpret the latter. Many statistical tests ultimately generate a 'p-value': the lower the p-value, the less likely it is that the null hypothesis is correct. A p-value of 0.03 indicates that if the two drugs are indeed equivalent, chance alone would be expected to yield the observed results three out of every 100 times one conducted the study. Conventionally, a p-value of  $< 0.05$  is taken to represent 'statistical significance', although this value can and should be adjusted in certain circumstances, for example if multiple comparisons are made [20, 21]. Confidence intervals can be used as an alternative to testing the null hypothesis [22, 23]; nonetheless, the conclusions reached by using confidence intervals are invariably the same. The real problem lies in how p-values are interpreted rather than how they are calculated [24]. An entirely different approach is to interpret a study's results mathematically in the context of prior knowledge (Bayesian statistics) [25, 26].

Regardless of the method of calculating or presenting statistical significance, the smaller the p-value, or the further away from zero the difference in confidence intervals, the less likely the result is to be a 'chance' finding and therefore the more likely it is that the difference between the two groups is indeed 'genuine'. But such a chance finding is still possible, albeit unlikely;

as Counsell et al. [27] point out, chance '...doesn't get the credit it deserves'. Furthermore, a low p-value does not exclude poor methodology in the conduct of the study.

*Power.* If the p-value in a drug study is, say, 0.07, does this mean there is genuinely no difference between the two drugs? Or does it mean that there might be a true difference, but that the study has simply failed to show it? It is specifically to help answer such questions that a power analysis is useful. The power analysis estimates how likely it is that a negative result can be 'believed'. One emerging problem is that some researchers (or their critics) are placing far too much emphasis upon power analysis [28]. One example demonstrates the type of misplaced faith in power analysis: '...at least 400 patients would be required to *prove* there is no statistically significant difference between the groups...' [29] (our emphasis). Such statements reveal a poor understanding of the scientific method and of the concept of scientific proof.

One reason for our concern is that the concept of power analysis itself has very serious limitations. Power analyses are only crude estimates of a sample size (indeed, the word 'crude' is emphasised by statisticians [30]). For example, two main elements that contribute to power for normally distributed continuous data are the difference between the means that is deemed important and the expected standard deviation (SD) of the measure of interest. Both of these are subject to serious shortcomings. The choice of what constitutes an 'important difference' is almost entirely subjective, and small but arbitrary adjustments to its value can have a great impact upon a study's calculated power. Where no previous data exist, the expected SD is usually taken from a pilot study, often without a control group, and by definition always smaller and less robust than the planned substantive study. In reality, the power analysis itself is probably best expressed in terms of a confidence interval: for example, 'Power analysis indicated that we would require 20–60 subjects to be 70–90% confident of detecting a difference between the means of 10–50 s' – although this is rarely done. The crudeness of the power analysis as a tool is reflected in the different sample sizes yielded by different methods. If we assume that for a hypothetical study, the important difference is 1.0 arbitrary units, with a standard deviation of 0.8 arbitrary units, then various calculations give a sample size per group (with 80% power at  $p < 0.05$ ) of 10 [30], 11 [31], 13 [32] and 18 [33]. So at best, power analysis only gives an approximate estimate of sample size. Indeed, Bacchetti [34] has suggested that a study's power should only be criticised if the study has no other shortcomings; in other words, all other features of a study

**Table 1** Ranges of sample sizes commonly used in different types of study. This represents a personal and superficial overview of the published literature in a wide variety of specialist journals.

Type of study	Examples of study question	Sample size
Laboratory or volunteer study in which conditions can be tightly controlled	a) Does drug X block Na <sup>+</sup> channels in nerve axons? b) The effect of hypoxia on heart rate in healthy humans	~ 5 to ~ 50
Clinical studies comparing 'common' outcomes	a) Is drug X better than drug Y for preventing nausea or vomiting after laparoscopy? b) Does device X function better than device Y for tracheal intubation?	~ 20 to ~ 200
Clinical studies comparing rare or serious outcomes for whole populations and/or when variability is great	a) Does technique X reduce mortality after surgery? b) Is drug X better than drug Y for hypertension?	~200 to ~10 000

(especially relating to its conduct) are far more important than its power analysis.

In practice, the actual sample sizes of studies are related to the type of outcome, the variability of the result and the statistical test used. We observe that in published studies, sample sizes tend to fall into three groups, though with considerable overlap (Table 1). Studies in which the outcome is easily and accurately measured, or there is little variability with few confounding factors (e.g. laboratory or volunteer studies in which experimental conditions can be tightly controlled with repeated testing) use smaller sample sizes. Variability usually increases once patients are involved, since conditions are harder to control. Clinical studies comparing 'common' anaesthetic outcomes have 'intermediate' sample sizes. Those in which the outcome is rare, variability is great and/or potential confounding factors are many, need much larger sample sizes [35]. It is striking how often studies and their sample sizes fall into these groups. It is tempting to speculate whether this phenomenon is a function of proper power analysis during the design of each study, whether investigators tend to use similar sample sizes because of the conventions adopted in their particular field of research, or whether they even 'massage' the power calculations to yield sample sizes that allow the study to be completed within a sensible timescale. There are ethical reasons for conducting power calculations as well as scientific ones [36], and we do not suggest that investigators should stop doing them, but given the crudeness of these calculations we cannot help but wonder whether the crudeness of Table 1 is any worse.

It can be seen from the above that retrospective studies, despite their attraction through not having to recruit subjects in advance – relying as they do on data already collected – are inherently weaker than prospective ones. Often, investigators have to rely on surrogate measures, recorded using tools that cannot be validated retrospectively, with little guarantee of blinding and uncertain definitions of the groups and their controls. The one

strength of retrospective studies is the relative ease with which very large samples can be studied.

### What makes a paper 'valuable'?

So much for a paper's components. We now turn to a different question: what makes one paper more 'valuable' than another, assuming proper care and attention to their conduct? To some extent, the answer to this question will always be subjective and depend upon the individual reader's own special interest, perspective and background. However, we suggest that there are some common elements that contribute to a paper's perceived 'value'. We divide these as being broadly related to the following:

- the type of question addressed and the answer provided;
- the strength of the evidence presented;
- certain 'decorative' aspects of a paper.

### The question addressed and the answers provided

The importance of the type of question relates to its relevance. This will depend on the perspective (e.g. specialty/subspecialty) of the person assessing the paper, the significance of the problem being addressed, and whether the correct and appropriate question is being asked. For clinical studies, the importance of a paper can depend on the magnitude of the change in practice likely to arise from the results, the speed with which that change might occur, and the feasibility of bringing about such a change. For example, the Collaborative Eclampsia Trial showed that magnesium sulphate prevented the recurrence of eclamptic convulsions and their complications more effectively and safely than alternative drugs [37]. The problem it addressed (treatment of eclampsia) is relevant to a wide range of practitioners and is both common and serious enough to make it a very significant one. Furthermore, the question (whether magnesium was more effective than alternatives) was clearly stated and appropriate for current knowledge at the time. The

change in clinical practice advocated by the study was small, cheap and easy to implement, yet was predicted to result in a huge reduction in maternal morbidity and mortality worldwide. Indeed, practice changed considerably and rapidly as a result [38].

A particular feature of the Collaborative Eclampsia Trial is that it provided a clear and unambiguous answer: magnesium is more effective than its alternatives and we should use it to prevent recurrence of convulsions. The Magpie Trial of magnesium for the prevention of eclampsia, however, whilst still in an area of great (possibly greater) clinical relevance and significance, raised as many questions as it answered: magnesium reduces the incidence of eclampsia but whether we should give it to all pre-eclamptics is less certain [39]. This example illustrates that sometimes the nature of a paper's conclusion can thus affect its 'value' as well as the nature of the question asked in the first place. Another way in which the answer provided may affect a paper's value relates to the amount of information given. For example, a recent study of the effects of spinal anaesthesia on respiratory function in obese parturients presented the median and interquartile ranges of the reduction in vital capacity and other variables, but not the minimum or maximum reductions [40]. In considering whether the reduction in vital capacity might be acceptable or dangerous for a particular patient, knowledge of the likely range of the changes (i.e. best and worst scenario) might be more relevant than knowing how the middle 50% of patients might be affected. While the value of the study is unaffected by this presentational detail, the value of the published paper is reduced as a result.

For studies in basic science, it is perhaps the effect on future thinking which is likely to be important, rather than any immediate practical application. For example, in the physical sciences, the theory of relativity solved no immediate practical problems, but clearly explained certain observations better than did more traditional theories [41]. An example from the biomedical sciences of papers that changed 'the way we think' is the discovery that DNA is a double helix [42]. Another is the finding that nitric oxide is fundamental to endothelial function; this discovery causes us to view in a new light all situations in which the response of the vasculature is involved [43, 44]. The scientists involved in these findings received Nobel Prizes, but there are many less well known examples. For instance, the discovery of a multiprotein complex (hypoxia-inducible factor, HIF) that binds specific DNA sequences termed 'hypoxia response elements', and that is central to erythropoietin production by the kidney [45], causes us to consider the possibility that this mechanism is a widespread, if not

universal, mechanism by which hypoxia may be detected at the cellular level [46].

It is possible (and indeed desirable) to reconcile clinical and scientific notions of 'value'. The manipulations of nitric oxide responses in sepsis are attempts to apply new understandings from basic science to solve a clinical problem [47, 48]. To refer back to the Collaborative Eclampsia Trial: a 'scientific' approach to its results would be to consider whether there is a fundamental reason why magnesium works better than other drugs, and that elucidating this reason will help us learn more about eclampsia as a disease process. However, we must acknowledge that (perhaps especially in anaesthetic related research) finding the relationship between basic and clinical science is frequently not easy. The relevance of basic science to clinicians' work is not always obvious to them, and scientists can often be frustrated by the inability to control tightly all the variables in clinical practice.

Another aspect of the type of question asked, which can apply equally to both clinical and basic science research, is its topicality and general profile – especially if the study's results can be condensed into an easily understood 'sound bite'. Certain topics are guaranteed to catch the eye more than others, and this can sometimes lead to the overriding of proper scientific considerations before publication by a journal. A recent example is the publication of Wakefield's [49] much-criticised study of the link between autism and vaccination (a highly topical issue) in the *Lancet* (a high impact factor journal – see below), and then the subsequent retraction of this paper [50].

In the wider scientific community, a very general rank order of the 'importance' of studies appears to have evolved, dependent in part upon the type of question addressed. This seems to have been adopted by the Research Assessment Exercise (RAE) with adverse implications for anaesthesia, and we discuss these issues further below.

### Strength of the evidence

The strength of the evidence presented in a paper depends, first, on those methodological aspects already mentioned and, second, on the greater weight afforded to certain types of study over others, a notion championed by Sackett and colleagues [51].

When assessing the first of these, the reader would do well to ask (as indeed the investigators themselves should have considered): 'Is there another possible explanation for these results other than the conclusions offered?' Tight methodology and attention to the above aspects may reduce or eliminate many of these alternative possibilities, but the more doubts that remain regarding hidden and overt biases, the less the weight that should be attached to

**Table 2** Classification of evidence levels (adapted from reference [52]). Level Ia is the 'strongest' evidence; level IV the 'weakest'.

Ia	Evidence obtained from meta-analysis of randomised controlled trials
Ib	Evidence obtained from at least one randomised controlled trial
IIa	Evidence obtained from at least one well-designed controlled study without randomization
IIb	Evidence obtained from at least one other type of well-designed quasi-experimental study
III	Evidence obtained from well-designed, non-experimental descriptive studies (such as comparative studies, correlation studies and case studies)
IV	Evidence obtained from expert committee reports or opinions and/or clinical experience of respected authorities

a set of results. This approach may seem an unduly negative initial stance but we prefer to think of it as a healthy state of 'septicaemia' [52].

Sackett and colleagues' now well-known proposal for a scale of the strength of evidence in biomedical studies ('evidence-based medicine') is outlined in Table 2. The emphasis on meta-analysis as a powerful tool has been much criticised [27, 53, 54]. Some have argued that meta-analysis is, in fact, really a form of collation-based medicine, rather than evidence-based, that really has no place in scientific studies seeking to test hypotheses [55]. The scheme mainly applies to interventional clinical trials, and not really to individual scientific (laboratory) experiments. These critics argue that fundamental hypotheses in science can only be addressed (in accordance with the scientific method, as described above) by a properly conducted experiment designed to test the prediction arising from the hypothesis, and not by simply combining the results of different experiments of varying quality. It is possible that meta-analysis is more useful for those studies concerning efficacy of a drug or treatment, or for *generating* hypotheses, rather than *answering* them [56]. Another problem with Sackett's scheme is that it relegates individual case reports and case series to the bottom of the hierarchy of evidence [57]. In many areas of medicine, especially in the more practical specialties such as anaesthesia, case reports can have a very persuasive effect on an individual's clinical practice, since they are based on direct clinical experience, and not on huge trials in which individual patients' experiences may be masked by grouping them into large samples and applying descriptive, summary statistics. If a patient receives a single drug and suffers anaphylaxis, for example, then that drug may cause anaphylaxis, whatever the results of a large clinical trial. Case reports are particularly valuable for very rare conditions, for very rare complications, or when they describe unexpected clinical benefits (or problems).

### 'Decorative' aspects of papers

We apply the above term to those aspects of a paper not relating strictly to its scientific content or context, but which nonetheless seem to have an unfortunate and disproportionate influence on the manner in which papers are received by the scientific or clinical community.

The institution where the work was carried out is one of these factors. For example, work from the universities of Oxford or Cambridge is likely to have greater impact than work of similar merit from, say, a rural sixth form college. The RAE can be criticised for having formalised this scheme: the currently successful institutes obtain better funding, and thus can undertake more research, which in turn is regarded as more influential since it emanates from the higher ranking institutes, and so on [58].

A related issue is the tendency for authors to 'pitch' their manuscript to the journal whose quality and circulation they feel adds to its importance. A journal's quality is artificially described by a single measure, the *impact factor* (an index of how often papers published therein are quoted by other authors). At the extremes, impact factors have some basis: a paper published in, say, a local parish gazette will have very limited readership, regardless of its scientific quality and so is unlikely to be widely quoted. Generally, though, impact factors are a poor measure of a journal's quality, and are open to manipulation by publishing strategy [59, 60]. However, some authors and institutions (and perhaps also the RAE) continue to judge a paper's worth largely by the journal in which it is published.

The reputation of the authors (or at least one of them) may also have an effect. A study challenging a current fundamental hypothesis is more likely to be read and accepted if a senior professor is an author. It is assumed, perhaps wrongly, that an established professor is unlikely to spend valuable time on unimportant work.

Finally, the funding of a study can influence its perceived importance. A study's support from a major funding body (e.g. Medical Research Council or Wellcome Trust) gives the impression that the work, or at least its preliminary design, has already undergone some peer review and that in a highly competitive process, it has been judged 'worthy' (we discuss this further below). The setting of priority areas for research by the National Health Service Research and Development strategy seems superficially reasonable, but may skew the influence of certain papers, since it is effectively Government policy that studies falling within the current priority areas are 'officially' more important. However, funding aspects can occasionally highlight conflicts of interest that reduce the perceived impact

of a paper (e.g. when funding has been obtained from tobacco companies to support a controversial paper relating to cancer research [61]). However, such conflicts may be hidden and remain undeclared.

### Measures of value used by grant-giving bodies

Thus far, we have discussed how an individual might undertake critical appraisal. On a wider level, grant-giving bodies have developed schemes by which they formally rank or score projects (using a process of peer-review) to inform their decisions about which research to fund. Although the scoring occurs before the study is conducted or published, the funding agencies feel that eventually the published paper(s) arising from the study will have a value corresponding to this score. Table 3 shows the Medical Research Council's published criteria by which it evaluates research applications [62]. It is clear that particular emphasis is given to projects that have a 'high impact on medical practice or scientific thinking', and to those that are 'very important in terms of disease burden or knowledge of mechanisms'.

The Higher Education Funding Council for England distributes about £1 billion of public funds each year to support research and research infrastructure, and the RAE is the method used to direct these funds to the universities producing the most 'valued' research.

The formulae used by RAE, although not made explicit, seem to place special emphasis on the ability of university departments to raise large independent income by way of grants (such as those awarded according to the criteria in Table 3), and to publish papers in high impact factor journals. In turn, universities distribute the money they receive to individual departments, using similar formulae [63]. Table 4 shows our (subjective) impression of the types of paper which seem to score more highly in this exercise.

These processes are open to much criticism [64], and whether they resemble good critical appraisal is open to debate. However, they remain *de facto* the means by which research is funded. If a specialty (such as anaesthesia) focuses on research which does not attain a high value in these processes, then these mechanisms will ensure (as they are designed to do) that the specialty attracts fewer funds. Eventually, as funding dries up, so will research posts and research opportunities, and then it may become impossible for that specialty to conduct *any* research at all. Many authors have described how anaesthesia is in such a crisis [65–69]. So while critical appraisal of papers by an individual (or by a specialty) might lead to one conclusion, appraisal by external organizations may come to a different conclusion. The dilemma for a specialty in this situation is to consider the extent to which it wishes to (or needs to) accept assessments of 'value' imposed from outside.

**Table 3** The Medical Research Council's published criteria for assessing research applications (adapted from [62]). In this scheme, we believe that even the most 'valued' anaesthetic-related publications will probably score a maximum of just 5 or 6, when assessed against research in other disciplines (but probably score of 9 or 10 if the comparisons are confined to those with other anaesthetic research).

Description	Score	Indicators
Excellent	10	Exceptional
	9	Is (or will be) be at the forefront internationally Addresses very important medical or scientific questions Likely to have a high impact on medical practice (or the relevant scientific field) At the leading edge internationally Very important in terms of disease burden or knowledge of mechanisms Likely to be very highly productive
Good quality	8	Intermediate
	7	Internationally competitive and at the forefront of UK work Highly productive and likely to have a significant impact on medical practice (if relevant) Very important in terms of disease burden or knowledge of mechanisms
Good quality	6	Intermediate
	5	At least nationally competitive Addresses reasonably important questions, and will be productive Good prospects of making some impact on medical practice (or the relevant scientific field) Any significant concerns about the research approach can be corrected easily
Potentially useful	4	Intermediate
	3	Plans contain some good ideas or opportunities but very unlikely to be productive or successful Major improvements would be needed to make the proposal competitive Fairly low expectation of success
Unacceptable	2	Intermediate
	1	Serious scientific or ethical concerns Should not be funded

1.	Papers testing (disproving) hypotheses that are widely applicable in different areas of research
2.	Papers applying scientific principles to practical (clinical) problems in novel ways (e.g. new technologies or treatments)
3.	Papers assessing clinical outcomes, particularly reduced mortality or serious morbidity
4.	Papers concerning improved efficiency/cost of service or concerning improved patient comfort
5.	Papers concerning technical notes, observations, equipment, measurement, etc.

**Table 4** Our impression of a rank order for the 'value' of experimental papers, as it seems to us to be understood by the wider scientific community. This impression is based upon criteria such as those presented in Tables 2 and 3. After category 5 in the table come all other types of non-experimental papers (editorials, reviews, case reports, etc.).

**Table 5** Steps for assessing a paper's value. How the paper performs at each step can promote or demote it from one imaginary 'in-tray' to another, such that it ends up in its final tray as a result of considering the various contributory aspects. Note that we have not specifically considered the *presentation* of the study (layout, use of graphs/tables, format and style, etc.), and we assume that the study is presented with sufficient clarity and in a logical and lucid manner.

**Conduct of the study**

<i>Clear hypothesis</i>	What question is being asked? Is there supporting evidence for the hypothesis? If the hypothesis were true, what results will be expected from the experiment?	
<i>Ethics</i>	Are there ethical issues?	
<i>Technical conduct</i>	Are surrogate measures being used? Are the measurement tools (both technological and assessment scores, etc.) appropriate? Is the study randomised properly? Is the study blinded as well as it could be? Are there well defined groups and proper controls? Are the statistics appropriate?	How is significance expressed and is it interpreted appropriately? Are confidence intervals used? Has sample size been considered (e.g. using a plausible power calculation)?
	Is the study retrospective or prospective?	
<b>Other factors</b>		
<i>The question addressed and the answer provided</i>	How relevant is the study?  How useful are the results?  Might the results change the way we <i>think</i> about certain problems or situations?	Which specialty/subspecialty is or might be affected? How significant is the problem addressed? What are the magnitude, speed, feasibility and impact of any likely change in practice? Does the study provide a clear answer or does it pose further questions?*
<i>Strength of evidence</i>	How persuasive is the evidence (e.g. the logic of argument; methods and statistics)? Are the results supported by evidence from other sources, or is the result unique? Could there be another explanation for these results? Is adequate information provided?	
<i>'Decorative aspects'**</i>	Who are the investigators and from which institution? What funding is there?	

\*Virtually all studies pose further questions but some, while providing useful information, do not directly answer the question they set out to address.  
\*\*One shouldn't be unduly influenced by these but they should be noted during the appraisal process.

**Conclusions**

**How does one assess a paper's value?**

We have discussed various aspects of a paper that can contribute to its value. We conclude here by suggesting

an approach to appraising a paper, based on the foregoing.

Consider a hypothetical stack of 'in-trays', with the top tray representing the most valuable papers and the bottom tray representing the least. The aim is to go

through each paper and at the end of the process, be able to place it in the appropriate tray. Each of the above factors – aspects of the study's conduct (including of course other aspects of its methodology not covered above), the nature of the question asked and the answers provided, the strength of the evidence and any persuasive decorative factors – should be considered at intervals when reading a paper, and each has the potential to move the paper up or down according to the impression gained thus far. The steps are summarised in Table 5. By this process the paper can be promoted or demoted many times before ending up in its final position, bearing in mind two basic principles. First, every paper will have some value, even if it ends up in the bottom tray (and any investigators who publish a paper should be commended for what is after all a considerable achievement, especially in the current regulatory climate). Second, different readers may place the same paper into different final trays – but this is a normal and appropriate aspect of judging value in any context. Discussion of the paper with colleagues (verbally or *via* the correspondence pages) should be directed towards sharing one's reasons for assigning the paper to a particular tray, and for being influenced in various directions (or not) by all aspects of the paper. This model makes it easy to demonstrate to others how the process of appraisal works, and is a good way of generating discussion and engaging others, which is a valuable process in itself.

### Going for gold?

We also wish, by our discussion, to highlight some wider considerations. One of our aims is to improve the value of all future work. On reading (and valuing) any given paper, the reader should ideally ask: 'How can we further improve on the approach used to answer the question posed?' The specific answer to this will depend, in part, on the shortcomings of the original paper (and we have indicated above where some of these might lie). However, in seeking to rectify these shortcomings by conducting a better study or experiment, the researcher will encounter the serious practical problems facing the specialty as a whole, especially those related to funding. Our discussion above highlights two important questions to consider. First, if funding is so critical to enable studies to be conducted, then we should not confine ourselves to considering what we alone value, but also ask: 'What do the major funding agencies value?' Second, within any given general topic which interests us as a specialty, there are always a range of possible questions we might address. It is pertinent here to consider: 'Which specific questions or hypotheses concerning this topic will increase the (perceived) value of our study?' It is possible that (even for

the topics of interest to our specialty) others place higher value on hypotheses different from those the specialty has traditionally considered important.

We hope that our discussion above might generate a debate which helps answer these two questions.

### Conflicts of interest (or, more accurately, reasons why the authors are interested in this topic)

S.Y. is an Editor of *Anaesthesia* and, through this role, interested in maintaining standards in all aspects of the conduct and publication of research. J.J.P is currently the Academic Strategy Officer of the Royal College of Anaesthetists, tasked with co-ordinating the Royal College's report on the future strategy for academic anaesthesia and anaesthetic research. Both have experience of basic science and clinical research, and have written papers that they wish could have been more valuable. The views above are the authors' own, and do not reflect policy of the Royal College of Anaesthetists, the Association of Anaesthetists of Great Britain and Ireland or the Editorial Board of *Anaesthesia*.

### References

- 1 Norman GR, Shannon SI. Effectiveness of instruction in critical appraisal (evidence-based medicine) skills: a critical appraisal. *Canadian Medical Association Journal* 1998; **158**: 177–81.
- 2 Souter MJ, Signorini DF. Meta-analysis: greater than the sum of its parts? *British Journal of Anaesthesia* 1997; **79**: 420–1.
- 3 Popper K. *The Logic of Scientific Discovery*. London: Hutchinson, 1959.
- 4 Kuhn TS. *The Structure of Scientific Revolutions*, 2nd edn. Chicago: University of Chicago Press, 1970.
- 5 Hall GM. *How to Write a Paper*. London: BMJ Books, 2003.
- 6 Zeiger M. Telling a clear story in a clinical anaesthesiology paper. *European Journal of Anaesthesiology* 1994; **11**: 313–20.
- 7 Yentis SM. Ethics again – hoops, loops and principles. *Anaesthesia* 2004; **59**: 316–7.
- 8 Echt DS, Liebson PR, Mitchell LB, *et al*. Mortality and morbidity in patients receiving encainide, flecainide, or placebo. The Cardiac Arrhythmia Suppression Trial. *New England Journal of Medicine* 1991; **324**: 781–8.
- 9 Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Annals of Internal Medicine* 1996; **125**: 605–13.
- 10 Robinson NP, Salmon P, Yentis SM. Maternal satisfaction. *International Journal of Obstetric Anesthesia* 1998; **7**: 32–7.
- 11 Rampil IJ. A primer for EEG signal processing in anaesthesia. *Anesthesiology* 1998; **89**: 980–1002.
- 12 Pandit JJ, Schmelze-Lubiecki B, Goodwin M, Saeed N. Bispectral index-guided management of anaesthesia in permanent vegetative state. *Anaesthesia* 2002; **57**: 1190–4.

- 13 Greif R, Akca O, Horn EP, Kurz A, Sessler DI. Supplemental perioperative oxygen to reduce the incidence of surgical-wound infection. *New England Journal of Medicine* 2000; **342**: 161–7.
- 14 Pryor KO, Fahey TJ 3rd, Lien CA, Goldstein PA. Surgical site infection and the routine use of perioperative hyperoxia in a general surgical population: a randomized controlled trial. *Journal of the American Medical Association* 2004; **291**: 79–87.
- 15 Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association* 1995; **273**: 408–12.
- 16 Smith I, Thwaites AJ. Target-controlled propofol vs. sevoflurane: a double-blind, randomised comparison in day-case anaesthesia. *Anaesthesia* 1999; **54**: 745–52.
- 17 Marfin AG, Pandit JJ, Hames KC, Papat MT, Yentis SM. Use of the bougie in simulated difficult intubation. 2. Comparison of single-use bougie with multiple-use bougie. *Anaesthesia* 2003; **58**: 852–5.
- 18 Chalmers I. Well informed uncertainties about the effects of treatments. *British Medical Journal* 2004; **328**: 475–6.
- 19 Altman DG. *Practical Statistics for Medical Research*. London: Chapman & Hall/CRC, 1999.
- 20 Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. *British Medical Journal* 1995; **310**: 170.
- 21 Bender R, Lange S. Multiple test procedures other than Bonferroni deserve wider use. *British Medical Journal* 1999; **318**: 600–1.
- 22 Bland M. Confidence intervals should be used in reporting trials. *British Medical Journal* 2000; **321**: 1351.
- 23 Altman DG, Machin D, Bryant TN, Gardner MJ. *Statistics with Confidence*, 2nd edn. London: BMJ Books, 2000.
- 24 Cobo E, Campbell MJ. P values are still useful. *British Medical Journal* 1994; **309**: 1439.
- 25 Bland JMG, Altman DG. Bayesians and frequentists. *British Medical Journal* 1998; **317**: 1151.
- 26 Freedman L. Bayesian statistical methods. A natural way to assess clinical evidence. *British Medical Journal* 1996; **313**: 569–70.
- 27 Counsell CE, Clarke MJ, Slattery J, Sandercock PA. The miracle of DICE therapy for acute stroke: fact or fictional product of subgroup analysis? *British Medical Journal* 1994; **309**: 1677–81.
- 28 Morris S. Underpowered and overbiased? Potentially unfair to the single-use bougie? *Anaesthesia* 2003; **58**: 1236–7.
- 29 Kendall JB, Russell GN, Scawn NDA, Akrofi M, Cowan CM, Fox MA. A prospective, randomised, single-blind pilot study to determine the effect of anaesthetic technique on troponin T release after off-pump coronary artery surgery. *Anaesthesia* 2004; **59**: 545–9.
- 30 Lehr R. Sixteen s squared over d squared: a relation for crude sample size estimates. *Statistics in Medicine* 1992; **11**: 1099–102.
- 31 Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *British Medical Journal* 1995; **311**: 1145–8.
- 32 Florey CV. Sample size for beginners. *British Medical Journal* 1993; **306**: 1181–4.
- 33 Young MJ, Bresnitz EA, Strom BL. Sample size nomograms for interpreting negative clinical studies. *Annals of Internal Medicine* 1983; **99**: 248–51.
- 34 Bacchetti P. Peer review of statistics in medical research: the other problem. *British Medical Journal* 2002; **324**: 1271–3.
- 35 Moore RA, Gavaghan D, Tramer MR, Collins SL, McQuay HJ. Size is everything – large amounts of information are needed to overcome random effects in estimating direction and magnitude of treatment effects. *Pain* 1998; **78**: 209–16.
- 36 Yentis SM. The struggle for power in anaesthetic studies. *Anaesthesia* 1996; **51**: 413–4.
- 37 The Eclampsia Trial Collaborative Group. Which anti-convulsant for women with eclampsia? Evidence from the Collaborative Eclampsia Trial. *Lancet* 1995; **345**: 1455–63.
- 38 Gulmezoglu AM, Duley L. Use of anticonvulsants in eclampsia and pre-eclampsia: survey of obstetricians in the United Kingdom and Republic of Ireland. *British Medical Journal* 1998; **316**: 975–6.
- 39 Yentis SM. The Magpie has landed: pre-eclampsia, magnesium sulphate and rational decisions. *International Journal of Obstetric Anaesthesia* 2002; **11**: 238–41.
- 40 von Ungern-Sternberg BS, Regli A, Bucher E, Reber A, Schneider MC. Impact of spinal anaesthesia and obesity on maternal respiratory function during elective Caesarean section. *Anaesthesia* 2004; **59**: 743–9.
- 41 Einstein A. *Relativity. The Special and General Theory*. London: Methuen & Co., 1916 (reprinted 2004).
- 42 Watson JD, Crick FHC. Molecular structure of nucleic acids. A structure for Deoxyribose Nucleic Acid. *Nature* 1953; **171**: 737.
- 43 Furchgott RF, Jothianandan D. Endothelium-dependent and -independent vasodilation involving cyclic GMP: relaxation induced by nitric oxide, carbon monoxide and light. *Blood Vessels* 1991; **28**: 52–61.
- 44 Ignarro LJ. Wei Lun Visiting Professorial Lecture: Nitric oxide in the regulation of vascular function: an historical overview. *Journal of Cardiac Surgery* 2002; **17**: 301–6.
- 45 Maxwell PH, Osmond MK, Pugh CW, *et al.* Identification of the renal erythropoietin-producing cells using transgenic mice. *Kidney International* 1993; **44**: 1149–62.
- 46 Pandit JJ, Maxwell PH. New insights into the regulation of erythropoietin production. *British Journal of Anaesthesia* 2000; **85**: 329–30.
- 47 Luiking YC, Poeze M, Dejong CH, Ramsay G, Deutz NE. Sepsis: an arginine deficiency state? *Critical Care Medicine* 2004; **32**: 2135–45.
- 48 Kilbourn R. Nitric oxide synthase inhibitors – a mechanism-based treatment of septic shock. *Critical Care Medicine* 1999; **27**: 857–8.

- 49 Wakefield AJ, Murch SH, Anthony A. Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *Lancet* 1998; **351**: 637–41.
- 50 Wakefield AJ. MMR – responding to retraction. *Lancet* 2004; **363**: 1327–8.
- 51 Cook DJ, Guyatt GH, Laupacis A, *et al.* Clinical recommendations using levels of evidence for antithrombotic agents. *Chest* 1995; **108**: 227S–230S.
- 52 Skrabanek P, McCormick J. *Follies and Fallacies in Medicine*. Glasgow: Taragon, 1989.
- 53 Feinstein AR. Meta-analysis. statistical alchemy for the 21st century. *Journal of Clinical Epidemiology* 1995; **48**: 71–9.
- 54 Egger M, Davey Smith G. Misleading meta-analysis. Lessons from 'an effective, safe, simple' intervention that wasn't. *British Medical Journal* 1995; **310**: 752–4.
- 55 Goodman NW. Knowledge is not measured by the tools of evidence-based medicine. *Anaesthesia* 2002; **57**: 1041.
- 56 Pandit JJ. The role of evidence-based methods in scientific study. *Anaesthesia* 2003; **58**: 184.
- 57 Mason RA. The case report – an endangered species? *Anaesthesia* 2001; **56**: 99–102.
- 58 Feldman S. Anaesthesia and the Research Assessment Exercise. *Anaesthesia* 1997; **52**: 1015–6.
- 59 Boldt J, Haisch G, Maleck WH. Changes in the impact factor of anaesthesia/critical care journals within the past 10 years. *Acta Anaesthesiologica Scandinavica* 2000; **44**: 842–9.
- 60 Hunter JM. The latest changes...and no more shorts. *British Journal of Anaesthesia* 2004; **92**: 7.
- 61 Enstrom JE, Kabat GC. Environmental tobacco smoke and tobacco related mortality in a prospective study of Californians, 1960–98. *British Medical Journal* 2003; **326**: 1057.
- 62 Medical Research Council Information on scoring systems [WWW document]. [http://www.mrc.ac.uk/index/funding/funding-specific\\_schemes/funding-current\\_grant\\_schemes/funding-scoring\\_system.htm](http://www.mrc.ac.uk/index/funding/funding-specific_schemes/funding-current_grant_schemes/funding-scoring_system.htm) (accessed 27 November 2004).
- 63 Oxford University. The University's Resource Allocation Method (RAM) (updated for 2004–5). *Oxford University Gazette* 2004; **4708** (Suppl. 1).
- 64 Joint Funding Body. Review of research assessment. Report by Sir Gareth Roberts to the UK funding bodies; issued for consultation May 2003. [WWW document]. <http://www.ra-review.ac.uk/reports/roberts.asp> (accessed 8 December 2004).
- 65 Harmer M. Academic anaesthesia – alive and kicking? *Anaesthesia* 2002; ; **57**: 1153–4.
- 66 Jackson RGM, Stamford JA, Strunin L. The canary is dead. *Anaesthesia*, 2003; **58**: 911–2.
- 67 Stuart-Smith K. CPR for the canary. *Anaesthesia* 2004; **59**: 296–7.
- 68 Feneck RO. Canary needs help; send for Willie Sutton. *Anaesthesia* 2004; **59**: 616–7.
- 69 Reade M. Resuscitating academic anaesthesia – or trying to breathe life into a dead corpse? *Anaesthesia* 2002; **57**: 1214.