Person. individ. Diff. Vol. 13, No. 4, pp. 393–399, 1992
Printed in Great Britain

0191-8869/92 $5.00 + 0.00
Pergamon Press plc

EDITORIAL

PEER REVIEW: ADVICE TO REFEREES AND CONTRIBUTORS

H. J. EYSENCK and S. B. G. EYSENCK
Department of Psychology, Institute of Psychiatry, University of London, De Crespigny Park, Denmark Hill, London SE5 8AF, England

(Received 25 November 1991)

In a recent article on editorial practices in psychology journals, Hartley (1988) found almost universal reliance on some form of peer review system, although he admitted that there were valid arguments against refereeing, e.g.

(1) Refereeing is unreliable—referees don't agree with each other.
(2) Refereeing is *not valid*, in that referees' judgments are not related to subsequent citation rates.
(3) Refereeing is *biased*, in that referees' judgments are coloured by the prestige of the author and the author's contribution.
(4) Refereeing is *inefficient*; it delays publication, it inhibits the publication of new ideas, and it is a time-consuming and costly business.
(5) Refereeing can be personally damaging, in that people, especially new authors, may find the experience painful and distressing.

Cummings and Frost (1985) and Lock (1986) have reviewed the evidence in a broad context, and have found support for all these criticisms. On the other hand, the evidence also suggests that (1) refereeing weeds out poor quality articles, and (2) improves both the content and the presentation of those remaining. Eysenck (1989), in reply, argued for the mixture of editorial practices which characterized the policy of "*Personality and Individual Differences*", pointing out its high degree of success when its "impact factor" in the Citation Index (0.89) is compared with other journals in the same vein, such as the *Journal of Personality* (0.60), *The Journal of Personality Assessment* (0.51), *The Journal of Research in Personality* (0.48), and many others using traditional methods of peer review.

The American Psychologist has published a whole series of papers and letters on the subject, dealing with the reliability of the refereeing process (e.g. Gottfredson, 1978; Scarr & Weber, 1978; Cichetti, 1980), the qualifications of editors (e.g. Lindsey, 1976, 1977), author reactions to the revision process (e.g. Cowen, Spinell, Hightower & Lotyczewski, 1987), citation impact and acceptance rate (e.g. Buffardi & Nichols, 1981), and the impact of value systems on the refereeing process (e.g. Ceci, Peters & Plotkin, 1985; Wong, 1981). It is the case that less than 2% of all articles that appear in APA journals are accepted in their original form; thus virtually all articles sent to APA journals require extensive revision (Eichorn & Vanden Bos, 1985). This seems an interesting comment on the ability of leading American psychologists to write acceptable articles for APA journals. The implication involved would only be acceptable if the reliability and validity of the reviewing process were a good deal higher than the evidence suggests. Editors of relevant journals are clearly doubtful about present-day reviewing processes.

The newly-formed *Journal of Social Behavior and Personality*, in its promotion literature, quotes a number of criticisms of current journal practices. Thus John Ziman, editor of *Science Progress*, writes that "All that we can be sure of is that present (editorial) practices are deeply flawed". Garth J. Thomas, former editor of the *Journal of Comparative and Physiological Psychology*, points out that "anything really novel is likely to be given a hard time in the publication process". Sol Tax, founder of *Current Anthropology*, states that "a bias that prevents competent work from entering

the arena of community scrutiny is much more damaging to individuals, institutions, and the scientific community than a bias that lets mediocre work slip through". L. R. Pondi, past associate editor of the *Administrative Sciences Quarterly*, points out that "our present corps of reviewers have been trained and conditioned in a persecution mentality—through poor treatment of their papers by other reviewers". L. D. Goodstein, past editor of the *Journal of Applied Behavioral Science*, argues that "it should come as no surprise that the interjudge agreement of these decisions is low, more or less bordering on chance". D. B. Rankin, co-ordinating editor of the *Journal of the American Statistical Association*, concurs: "All who routinely submit articles for publication realize the Monte Carlo nature of review (random selection)." Finally, a study by Bradley (1981) is quoted in which professors were asked to comment on the peer review of their most recently published article. Seventy-six per cent encountered pressure to conform to the strictly subjective preferences of the reviewers; 67% encountered inferior expertise; 60% encountered concentration on trivia; 40% encountered careless reading by referees. (These are comments by authors whose work was accepted, not rejected) No wonder that R. S. Yalow, Nobel Laureate in Physiology/Medicine, commented: "I think that what has been demonstrated by this study is reviewer and editorial incompetence".

Clearly this is a complex and disputatious area in which the only clear fact is the unhappiness and disillusionment of most authors, faced with what amounts to a rejection rate of over 98% in APA journals, compared with something like 20% in journals of physics and astronomy. Worst of all is the fact that *peer* review turns out to be seldom done by one's *peers*, i.e. psychologists of equal status, publication and Citation Index rank, but by newly-fledged Ph.Ds. of little or no standing, more concerned to find and exaggerate minor flaws (often imaginary) rather than evaluate the article as a whole, and judge its contribution to science.

Most criticism has been aimed at the low *reliability* (inter-referee agreement) of judgments in this field. Thus, for instance, Bowen, Perloff and Jacoby (1977) reported a coefficient of concordance (Kendall's W) of 0.11 in judgments concerning the selection of the best paper in a contest in which 10 of the past presidents of the Division concerned ranked the papers submitted; in such a competition there is of course a restriction of range, but the whole exercise must be futile if past presidents only share 1% of agreement in common

McReynolds (1971) reported correlations of 0.45 among rating of papers submitted for presentation at a national convention, but this is probably higher than would be found in actual journal reviewing processes because all the papers were reviewed simultaneously. Scott (1974) reported a study more closely approximating actual reviewing processes, and found measures of inter-referee agreement ranging from 0.07 to 0.37 for attributes of the manuscripts; for recommendations to the editor of the *Journal of Personality and Social Psychology* the correlation between two referees was 0.26, i.e. less than 7% overlap

Gottfredson (1978) reported figures of 0.41 agreement on quality and 0.35 on impact, thus achieving a somewhat more acceptable agreement level of between 12 and 17%. Even that, of course, is well below any acceptable standard. The author developed 9 scales, using a factor analytic approach, for judging papers; inter-referee reliabilities for these scales ranged from 0.16 (do nots), 0.19 (magnitude of problem/interest) and 0.20 (stylistic/compositional dos) to more acceptable 0.50 (substantive dos), 0.49 (data grinders) and 0.45 (where do we go from here?). The average was very low for these 9 criteria (0.33), i.e. just above 10% agreement.

Low reliability precludes high validity, but it would be interesting to compare referee judgments with some extraneous estimate of quality and impact. Gottfredson (1978) used Citation Index measures as his criterion, perhaps the best available, and found correlations of 0.20 with referee's judgments of quality, and 0.27 with referee's judgments of impact. These very low correlations were higher for high-citation articles, totally absent for low-citation articles; clearly low-impact articles are not readily recognized by referees.

These are just some representative studies of a large literature; data are inconsistent but suggest on the whole that under routine refereeing conditions reliability of judgments tends to be low, and validity even lower. Some authors have suggested ways of improving the situation (e.g. Brackhill & Korton, 1976; Wolff, 1970, 1973), but it is not obvious that their suggestions have been widely followed, or have provided any improvement in a very unsatisfactory situation.

One major problem in this whole field, which bedevils discussion, is the assumption that decisions of excellence can be made along one dimension of good or bad. Nothing could be further from the truth. In nearly all fields judgments of good and bad are multi-dimensional, compounded of unrelated factors. It may be useful to list some of the relevant variables which together constitute the multi-dimensional universe within which judgments have to be made.

1. Overall contributions to science vs ho-hum ordinariness

Some articles clearly do little but repeat along much-trodden paths, things demonstrated many times before. There are bandwaggon themes, such as "Type A–Type B" research, which seldom threaten to set the Thames on fire. On the other hand, there are occasional articles which point out new vistas, suggest new horizons, and obviously add something to our knowledge. This factor overlaps with some of the others considered below, but cannot be altogether reduced to their sum.

2. Novelty vs ordinary science

Really new ideas have always had a particularly difficult life in fighting for recognition. When R. Fisher submitted his revolutionary 1918 paper on generalizing Mendel's laws to polygenetic inheritance, it was rejected by The Royal Society, the referees being, as he said rather spitefully, "a statistician who knew no genetics, and a geneticist who knew no statistics" (Box, 1978). This absolutely fundamental paper, extending the analyses of genetic factors to a whole new universe, was too novel to be acceptable to lower mortals. Many other examples can be found in the history of science. Practitioners of ordinary science usually have a much easier time in gaining acceptance of their papers.

3. Content breadth and narrowness

The content of the paper may cover a wide field, or it may deal with just a very narrow aspect of the field. Inevitably the broad approach is of greater interest and value, but it is also likely to be less sound methodologically, and less definite in its implications. Yet it may make a much bigger contribution to the progress of science, just because of the many implications of its findings.

4. Theory-oriented vs heuristic

Studies may derive from a general theory which is being tested, or they may simply report the finding of correlations between certain variables, correlations not particularly relevant to any specific theory. Most papers on offer seem to be of the second kind, although efforts are often made to make the finding appear to be relevant to one theory or another.

5. Experimental or correlational

Deductions from theory can of course be tested along correlational lines. Factor analysis can be used to suggest hypotheses or to *test* hypotheses. However, a better way of testing deductions is usually by actual experiment, whether psychological or physiological.

6. Rigour vs approximation

We all like rigour, or at least the sound of the word, but historically rigour correlates negatively with novelty. Newton's calculus was severely criticized because of the lack of rigour in its development, and it was not until Cauchy published his *Cours d'Analyse* 150 yr later that calculus became mathematically rigorous. Should peer review have kept Newton from publishing his Principia?

7. Replication vs independence

We all believe that results have to be replicated in order to be accepted, but we tend to undervalue such replication studies, and turn them down in favour of "something less repetitive". Yet replication is very desirable, and without it few results should gain general acceptance (Neuliep, 1990).

8. Positive vs negative results

Most journals shun negative results, and prefer to print positive outcomes. Yet according to Popper (1935) the basic principle of science is *falsification*, and of course the successs of falsification is the triumph of the null hypothesis. Should we not then welcome negative outcomes? Perhaps only when a sensible theory of wide acceptance is being tested, and the test is a rigorous one. Certainly each case should be judged on its own merits.

9. Power of statistical tests

As Cohen (1962), Iversky and Kuhneman (1971) and Rossi (1990) have pointed out, the statistical *power* of psychological research is generally very low, suggesting that many published articles may inflate the occurrence of type I error rates through insufficient power. This is a serious issue, usually disregarded by contributor and referees alike; it deserves much greater attention when considering acceptance or rejection.

10. Short paper vs long?

There is an obvious advantage to a short paper over a long one, in that the editor can squeeze more of them into the space at his disposal. But important papers may need a detailed theoretical discussion, and lengthy presentation of methods and results. Perhaps we should base our judgment on whether the length of the paper is appropriate to its contents; too much compression or too much padding may both be objectionable.

11. Simple vs elaborate statistics

Complex statistics may overwhelm the easily influenced tyro, and simple statistics may leave out necessary controls; it is often difficult to judge the appropriate degree of sophistication. Where possible simplicity is preferable, as long as nothing important is lost in the process.

12. Style of writing

One is inevitably influenced in one's judgment by the writer's style; equally many reviewers try to rewrite the paper in their own image (as well as making the author test theories they happen to be interested in, or using tests they themselves would have preferred). One should remember that the paper is the writer's work, not the reviewer's, and that changes should only be made to correct actual grammatical errors, faulty punctuation, or utterly unintelligible sentence structure.

13. Sample—students vs non-students

American research has homed in on students, mentally disturbed people and rats to the exclusion almost of ordinary, normal persons. The make-up of the sample studied should certainly weigh with us in deciding about the fate of a paper—restrictions to unrepresentative groups may or may not count against a submission, depending on the relevance of generalizability.

14. National vs international

Personality and Individual Differences (PAID) is an *international* journal, while APA and BPS journals are clearly national in character, although the occasional contribution from abroad slips through. *PAID* emphasizes and solicits contributions from all countries, and would give some slight preference to submissions from countries whose contributions to psychological research have not been numerous in the past—provided of course that their quality is up to the mark. We have gone out of our way to encourage such contributions, clean up the English where needed, and generally facilitate acceptance of papers from such sources.

These 14 areas of judgment largely duplicate and extend those isolated by Gottfredson (1978) by means of factor analysis. Component (1) is a list of "do nots", i.e. practices to avoid. Components (2) and (3) seem to suggest a differentiation of two types of "dos"—those dealing primarily with scientific or substantial matters, and those dealing with stylistic, compositional or expository matters. Component (4) suggests the importance of originality, and component (5) might be labelled "trivial". Component (6) refers to scientific advancement, while (7) refers to "data grinders" or "brute empiricists". Component (8) is labelled "routine" or "ho-hum" research, and

the final component refers to narrowness of research concerns. Some of our areas are not covered, partly because they are of less concern to journals published by a national association.

Here then are 14 mainly independent areas of judgment. Not only is judgment within each area probably quite unreliable (although probably less so than any overall judgment), but the stress a reviewer lays on each element is probably different from the stress laid on it by other reviewers, or the editor. It is the complexity of these relations which makes decision so hard; it is seldom a question of a paper being better than another, but rather whether we should print a very original paper, less rigorous methodologically but theory-oriented and experimental, or a relatively unoriginal one with great rigour and excellent statistical treatment.

One obvious solution—print them all, unless they are obviously too bad, uninteresting, or non-rigorous to pass—is clearly impossible because of the contingencies of space. There are only so many pages per year allowed by the publisher, chasing inexorable economic laws; some elimination there has to be, and the process of judgment cannot be avoided. For a time we can procrastinate, at the expense of the waiting time from acceptance to publishing getting longer and longer, but finally even this process catches up with the editor, and he cannot avoid cutting down any longer. We have tried to keep this waiting down to as near the minimum of 6 months as possible, and not accept papers at a rate which would bring this period up to the 12–18 months or so typical of many journals. So what is our advice to referees and contributors alike? As the number of contributions increases, while the number of pages remains fixed around the 1200 mark, more and more good, workmanlike articles which we would have been glad to publish will have to be turned down in favour of others judged "better" in terms of one or other, or a combination, of the criteria listed above. In this choice, many errors will inevitably occur, if the literature quoted is any guide.

In the past year we published almost 200 papers; if each had been refereed by 2 judges that would amount to 400 reports, not taking into account rejected papers. If we agree that no referee should be asked to referee more than 5 papers per annum (which is quite a burden, considering that such referees are likely to be asked to act for other journals as well), then we would have to find almost 100 judges for our journal alone. Can it be seriously argued that there are 100 psychologists in Great Britain who could be called the "peers" of our contributors, or who just have enough knowledge of the whole field of personality to form a meaningful judgment? Of course we do use referees from abroad, particularly the U.S.A. when required, but these also have many similar calls on their time, and must constitute an exceptional resource. A similar logistic problem faces most journals, depressing the qualifications of "peer" judges, particularly when new and unusual contributions are to be judged.

What conclusion does the argument suggest? It suggests that the major function of referees should be to point out actual defects (theoretical, methodological, statistical, instrumental, or with respect to the conclusions drawn), leaving it to the editor(s) to decide whether these defects are crucial and fundamental, leading to rejection, whether they can be abolished by rewriting or reworking, or whether they are unimportant blemishes which do not detract from the value of the research. Some referees get irate about the use of certain coefficients of correlation (phi or tetrachoric), or certain methods of factor rotation, or decisions about numbers of factor retained, matters which are largely subjective and not for the most part good reasons for rejection.

Referees can of course also usefully point our special virtues of the paper, in any of the fields just mentioned, and editors will of course take such comments into account. What we think editors should *not* do, but often insist on doing, is to abrogate the right of decision, and implicitly follow the judgment of the referees. Even worse, they may follow the judgment of one referee, who advocates rejection, and go against the advice of other referees who advise acceptance. Advice to reject may be based on the discovery of latent faults in the MS, but it may also be based on latent bias in the judge, miscomprehension, and failure to see the many good points compensating for a few bad ones. Some judges seek perfection in papers submitted, and refuse to compromise with reality. Their advice should not be followed slavishly.

To say this does not devalue the function of the referee. Discovery of faults and errors is vital; such decisions can be more objective than judgments concerning the overall contribution of the submission. Our recommendation emphasizes the role of the editor(s), and insists on the importance of their input. After all, when referee reliability lies somewhere between 0.2 and 0.4,

they cannot be regarded as infallible Neither should their voice be disregarded; there is a meaningful compromise which a good editor will strive to achieve. In too many cases does a negative report act to blackball a submission, even though other referees may approve of acceptance. No one person should have the power of veto, and much of the opposition to the peer review system stems from this abrogation by editors of their right to reject referees' advice.

What can we say to potential contributors who wish to improve their chances of having their papers accepted? Given the low reliability and validity of referees' judgments, we cannot promise to accept all worthy papers, and keep out all unworthy ones. Considering the high quality of papers submitted, it will in future be necessary to reject some papers which in the past we would have published quite gladly, simply because even better papers are available to cover our quota of papers—page allocation sets us a limit which could only be exceeded if we were willing to lengthen the waiting period between acceptance and publication, and even then only for a limited period—a path we are unwilling to travel.

Some points are obvious. We prefer papers which solidly advance our understanding of personality and individual differences; papers which are novel and original in approach, theory or application; papers which have broad reference and implications; papers which derive from established theory, and test deductions therefrom; papers which are experimental rather than simply correlational. We look for rigour, but not rigid application of arbitrary statistical principles, we value replication of important research, whether results are positive or negative. We frown on papers which are self-indulgent as far as length is concerned, but are not put off by lengthy papers where content justifies length. We are not impressed by statistical over-elaboration, but require of course adequate analysis of data. We prefer samples of ordinary people, but do not of course reject automatically samples of sophomores, or rats. We welcome contributions from countries not in the mainstream of scientific advance, as far as psychology is concerned, and do not presume to correct a writer's style unless what he says is ungrammatical, unclear or unidiomatic.

We are unhappy with "ho-hum" data-grubbing and "data-grinding", the endless analysis without a theoretical purpose of correlations between different scales or inventories; narrow research concerns and works not related to theory testing. Most of these concerns would probably be shared by our colleagues who edit the *Journal of Personality*, the *Journal of Research in Personality and Social Psychology* or the *Journal of Personality and Social Psychology;* if *Personality and Individual Differences* has a taste of its own, this is probably more a matter of past history than of present-day differentiation. The type of research we were publishing in the early-1980s was not then popular, or readily acceptable in, say, the *Journal of Personality and Social Psychology*. This has now changed dramatically, and the contents of these journals are probably largely interchangeable. Some differences still remain, but they are probably minor, and likely to disappear.

Do we then have an answer to the problems of peer review, viz. low reliability and low validity? In principle the answer must be in the negative; like democracy, which is the least bad form of government, so peer review may be the least bad form of selection. We tend to rely on a rather small but highly selected group of referees, rather than a large band of perhaps less experienced judges. When we have to go outside this group for specialized expertise, we try to get the best advice available. We try to give reality to editorial responsibility in deciding between contrary referees' opinions. But overall the problem can have no easy solution, and we felt it was only fair to share our misgivings with potential contributors, even though no ideal solution appears to be in prospect.

REFERENCES

Bowen, D. D., Perloff, R. & Jacoby, J. (1977). Improving manuscript evaluation procedures. *American Psychologist, 27*, 221–225.

Box, J. F. (1978). *R. A. Fisher: The Life of a Scientist*. New York: John Wiley.

Brackbill, Y. & Korton, F. (1976). Journal reviewing practices: authors' and APA members' suggestions for revision. *American Psychologist, 31*, 675–678.

Bradley, (1981). Quoted by J. E. Lloyd (1985).

Buffardi, L. C. & Nichols, J. A. (1981). Citation impact, acceptance rate, and APA journals. *American Psychologist, 36*, 1453–1456.

Ceci, S. J., Peters, D. P. & Plotkin, J. (1985). Human subjects review, personal values, and the regulation of social science research. *American Psychologist, 40*, 994–995.

Cichetti, D. V. (1980). Reliability of reviews for the *American Psychologist*: a biostatistical assessment of the data. *American Psychologist, 35,* 300–304.
Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal & Social Psychology, 65,* 145–153.
Cowen, E. L., Spinell, A., Hightower, D. & Lotyczewski, B. S. (1987). Author reactions to the manuscript revision process. *American Psychologist, 42,* 403–405.
Cummings, L. L. & Frost, P. J. (Eds) (1985). *Publishing in the Organisational Sciences.* Hamwood, IL: Irwin Inc.
Eichorn, D. H. & Vanden Bos, G. R. (1985). Dissemination of scientific and professional knowledge: journal publication within APA journals. *American Psychologist, 40,* 1309–1316.
Eysenck, H. J. (1989). Refereeing in psychology journals. *The Psychologist, 3,* 98–99.
Gottfredson, S. D. (1978). Evaluating psychological research reports: dimensions, reliability and correlates of quality judgments. *American Psychologist, 33,* 920–934.
Hartley, J. (1988). Editorial practices in psychology journals. *The Psychologist, 1,* 428–430.
Iversky, A. & Kuhneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.
Lindsey, D. (1976). Distinction, achievement and editorial board membership. *American Psychologist, 31,* 799–804.
Lindsey, D. (1977). How qualified are editors? *American Psychologist, 32,* 578–585.
Lloyd, J. E. (1985). Selling scholarship down the river: the pernicious aspects of peer review. *Chronicles of Higher Education, 14, June 26.* p. 64.
Lock, S. (1986). *A Difficult Balance: Editorial Peer Review in Medicine.* Philadelphia: ISI Press.
McReynolds, P. (1971). Reliability of ratings of research papers. *American Psychologist, 26,* 400–401.
Neuliep, J. W. (Ed.) (1990). Handbook of replication research in the behavioral and social sciences. *Journal of Social Behavior & Personality, 5,* No. 4.
Popper, K. (1935). *Logik der Forschung.* Wien: Springer.
Rossi, J. S. (1990). Statistical power of psychological research: what have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58,* 646–656.
Scarr, S. & Weber, B. L. R. (1978). The reliability of reviews for the American Psychologist. *American Psychologist, 33,* 935.
Scott, W. A. (1974). Inter-referee agreement on some characteristics of manuscripts submitted to the Journal of Personality and Social Psychology. *American Psychologist, 29,* 698–702.
Trersky, A. & Kuhneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76,* 105–110.
Wolff, W. M. (1970). A study of criteria for journal manuscripts. *American Psychologist, 75,* 636–639.
Wolff, W. M. (1973). Publication problems in psychology and an explicit evaluation schema for manuscripts. *American Psychologist, 28,* 257–261.
Wong, P. T. (1981). Implicit editorial policies and the integrity of psychology as an empirical science. *American Psychologist, 36,* 690–691.