Psychology Teaching Review (1994) 3, 1

Evidence of sex bias knocked out by blunt instrument? A comment on Hartley (1992)

Clare Bradley, Department of Psychology, Royal Holloway, University of London

This critique focuses on the methodology used in studies reported in a paper in a previous issue of this journal (Hartley, 1992). Hartley concluded that his studies showed no evidence of the sex bias in examining demonstrated in earlier work by Bradley (1984). It is here suggested that Hartley's studies were not designed in a manner that might be expected to demonstrate any kind of sex bias. It is argued that the "blunter but more practical measure" used by Hartley was unsuited to the task of detecting sex bias and the lack of evidence for sex bias in his studies is no reason to conclude that sex bias did not operate in the department studied.

In the second issue of this journal James Hartley reported four studies which he suggested attempted to replicate and extend my earlier studies of sex bias in evaluation of students (Bradley, 1984). The first three of Hartley's studies used what he described as a "blunter but more practical measure" than I had used. In the first of his studies, Hartley investigated the marking of third-year undergraduate projects in a psychology department and reported findings "which do not replicate those of Bradley". Although it is clearly the case that Hartley's findings did not replicate those of my 1984 study, the nature of the study he devised would not have permitted such a replication. Hartley tested for a different pattern of sex bias using a methodology that appeared to have been chosen for its convenience, despite its being unlikely to detect any form of sex bias and despite being entirely unsuitable for the detection of the kind of two-way sex bias demonstrated in my 1984 study.

I had predicted from the literature on sex bias in evaluation, notably a review by Nieva and Gutek (1980), that sex bias in marking of student projects would be manifested more by second markers than by first markers who had supervised the student and who usually knew both the student and the research area better than the second marker. The two-way bias that I predicted and found was such that second markers marked men students more extremely, while marking women students towards the lower second class relative to the first markers. Thus, relative to first markers, second markers favoured men students and marked women students lower at the upper end of the marking distribution, while favouring women students and marking men students down at the lower end. Investigation of predicted patterns of differences between marks awarded by two markers, where there was reason to expect one marker (the supervisor) to be less prone to sex bias than the other (the second marker), allowed for the investigation of sex bias independently of any actual differences in achievement of men and women students.

The four university departments which contributed data for my study were three social science and one design department.

(They were not, as Hartley assumed, all psychology departments.) In one of the social science departments, first and second markers' marks went forward as two separate marks to the final examiners' meeting. Thus bias shown by second markers in that department would be directly reflected in the profiles of marks which determined final degree class awarded. In the other three departments, the two markers discussed their initial independent marks with a view to reaching an agreed mark. For these three departments, I examined the pairs of initial marks which differed across a class boundary, to determine which marker's mark more closely approximated to the final agreed mark. There was no significant difference. Thus, it was not the case that bias from the second marker would have been eradicated by any greater influence of the first marker. Nor was it the case that the more bias-prone second markers demonstrated greater influence over the first markers. Both markers were equally influential in resolving initial disagreements and determining the agreed mark.

In Hartley's studies, he did not have information about which marker was the supervisor and which the second marker. He was, therefore, unable to test predictions about the nature of differences between first and second markers. Instead, he took the remarkably few cases where first and second markers' marks differed across a class boundary and looked to see whether the agreed mark corresponded to the higher or lower of the two marks for men and for women students separately, regardless of which class boundary the marks straddled. It seems that Hartley, was testing the hypothesis that the marks would be resolved upwards for men and downwards for women regardless of the level of performance of the student. For such a hypothesis to be supported it would require that, when marks differed across a class boundary (for whatever reason), both markers (supervisor as well as second marker) agree on the higher mark for men and on the lower mark for women. Although there are instances of this kind of simple pro-male/anti-female bias in the broader literature on evaluation bias, it was not the pattern of bias that I predicted, tested for or found in my 1984 study. Indeed, in my 1984 study I found no differences between markers in their influence in determining the agreed mark. There was an absence of any tendency to resolve differences in the predicted direction of bias in the four university departments I studied too. The sex bias demonstrated was manifested at the earlier stage where the initial difference between marks was generated, not when that difference was resolved. Thus Hartley found "no real evidence to support the idea that sex bias occurs in the marking of student projects ... in this particular university department", but he had not looked for the evidence where previous research had shown it to be found.

In his second and third studies, Hartley examined disagreements over exam scripts instead of projects. Again he looked only at disagreements apparently to test for one-way pro-male/anti-female bias in resolving those differences. The direction of bias hypothesized was nowhere stated. Although the literature provides no basis for anticipating pro-female/anti-male bias under such circumstances as these and a one-tailed test for pro-male/anti-female bias could be justified a low power two-tailed test was the test that Hartley chose to use. The same methodology was applied to exam scripts without, apparently, considering that bias shown by markers of exam scripts may differ from that shown by project markers. With the projects, there was reason to believe that bias, if it occurred, would be reflected in disagreements between supervisor and second marker in circumstances where the supervisor had more information about the candidate's work and was more knowledgeable about the research area than was the second marker (Nieva and Gutek, 1980). Where the marking of exams was concerned, there was no clear reason for expecting sex bias to be reflected in disagreements between markers. Indeed, where both markers are equally prone to bias and share the same cultural stereotypes, one of the effects of sex bias is, artificially, to increase the number of agreements about the class of mark to be awarded and identical marks from two markers may be equally biased.

Where marking of named exam scripts is concerned, the literature on sex bias in evaluation suggests that the nature of any bias will depend on the knowledge and experience of the particular markers of the particular topic areas involved as well as their knowledge of the individual candidates (Nieva

and Gutek, 1980). It may well be that for some exam papers, the first marker has taught the course while the second is much less familiar with the topics covered. In such cases we might predict that the first marker would be less prone to bias than the second and the nature of discrepancies between marks for men and women at the various class boundaries could usefully be investigated. For other exam papers where two members of staff share the teaching and examining of a course and share the same cultural stereotypes, the markers may be equally prone to bias on the paper taken as a whole, but on any one particular question, the marker who set the question may be less bias-prone. A range of strategies are available for resolving any disagreements in initial marks awarded and include strategies that will preserve at least some of the effects of sex bias (e.g. taking the average of the two marks) and strategies that minimize the effects of sex bias (e.g. acknowledging the question setter's better-informed judgement and placing greater weight on the question setter's mark). The most serious instances of sex bias, and those least susceptible to detection, will be likely to occur when two markers with a similar grasp of the topic areas covered, and holding similar cultural stereotypes, mark, under considerable time pressure, named scripts with essay answers from large numbers of students with whom they have had limited contact. In such instances we may well see relatively few disagreements between the markers. Hartley appeared not to consider the differences between project marking and exam marking nor the many different circumstances of exam marking. In particular, he overlooked the possibility that sex bias can manifest as consensus when exam script markers are equally prone to bias. Hartley's blunt instrument was applied even less appropriately to the task of detecting sex bias in the marking of exams than it had been to the task of detecting sex bias in the marking of projects.

Hartley increased the likelihood that consensus would disguise sex bias by excluding from his analyses, marks associated with the one woman member of the department he studied. His reasons for excluding the woman examiner's marks were not specified. I assume he thought that a woman examiner would be less likely to show pro-male/anti-female bias. However, if so, in excluding the marker who might be thought least likely to share the cultural stereotypes of the male majority, Hartley may also be excluding from the analyses those very disagreements which, given the methodology he has chosen, are the ones most likely to reflect sex bias, particularly where the exam scripts are concerned and first and second markers are equally prone to bias. In the absence of any basis for assuming one marker to be more prone to sex bias than the other in each pair of examination script markers, Hartley's second and third studies cloud the issue rather than extend the research and can offer no evidence of relevance to the question of sex bias in the marking of exams.

Hartley's fourth and final study addressed the question of whether examiners can tell the sex of candidates' handwriting, thereby reducing the value of blind marking as a means of eliminating sex bias. This issue was not addressed in my earlier research on sex bias in project marking where the projects involved were not handwritten but typed. Hartley's data contributed further evidence to suggest that handwriting, together perhaps with other characteristics of the script (which were not controlled for), may provide cues to candidate sex which may in turn influence the marks awarded. However, although seven of the 20 scripts (35 per cent) were identified by sex accurately by all ten of the male markers studied, the majority of the scripts were less readily categorized. In such cases where handwriting is a clear indicator of sex of author, sex bias might operate on unnamed scripts via impressions gleaned from handwriting. However, for 65 per cent of scripts there was less than perfect agreement. Though undoubtedly an imperfect solution, blind marking would seem nevertheless to have substantial scope for reduction of sex bias in the majority of cases where handwriting is not a clear indicator of author sex. At the same time, blind marking can be expected to reduce "halo" effects and other forms of bias. Though offering reasons in addition to reduction of sex bias in support of his case, Hartley also argued for the introduction of blind marking.

Hartley's first three studies were not only conceptually muddled but the quantity and representativeness of the data were inadequate: the sample sizes were small and

large amounts of data were only vaguely and partially accounted for. For example, Table 2 of Hartley's paper reported that in year 1, there were 11 cases of disagreements out of a total of 72 projects marked. Disagreements between markers in other departments studied have been closer to 50 per cent (Bradley, 1984; 1993) than the 15 per cent reported Hartley. The exclusion of one woman marker would not seem to account for some 25 missing marks. In describing the method of his second study, Hartley reported that not all markers supplied their data (p.91). Though not stated, it looks as though not all markers supplied their data for study 1 either. Markers who chose not to make such data available for scrutiny are likely to be more, rather than less, biased in their marking. Analyses conducted on only those marks that markers have been willing to make available are likely to underestimate the presence of bias even in the best designed of studies. In Hartley's studies the designs were unsuitable to the task of detecting sex bias, as well as the data being seriously limited in quantity and quality.

The conceptual muddle in Hartley's paper was not restricted to the design and analysis of his own studies. The introduction to the paper lent heavily on an unpublished report from the University of Wales College of Cardiff investigating final examinations marks in the Faculties of Humanities and Social Studies (Parry-Langdon, 1990). Hartley reproduced a table of data from that unpublished report which was said to demonstrate that there were no significant changes in the proportions of men and women getting "good degrees" awarded to men versus the proportion awarded to women before blind marking followed by a similar comparison after blind marking had been introduced. The proportion of men and women students involved were different in the two periods of analysis and it was suggested that the comparisons presented, in accounting for the change in proportions of men and women, were an improvement on the comparisons made in an earlier published report by Belsey (1988) which showed that, in the Department of English at the same University, women were awarded more "good degrees" after blind marking had been introduced than they were before. Hartley suggested that the earlier results of Belsey were confounded by changes in the proportions of males and females in the separate periods of analysis. However, Belsey made the entirely appropriate comparison of the percentage of women being awarded "good degrees" before marking with the percentage of "good degrees" awarded to women after blind marking. Percentages of men receiving "good degrees" before and after blind marking were determined separately such that the results would not be affected by any change in the proportions of men or women. Only the Parry-Langdon data shown in Hartley's Table 1, needed to take account of the proportions of men and women in interpreting their between-sex comparisons. Belsey's within-sex comparisons were quite appropriately interpreted in a straightforward manner, Thus the Parry-Langdon data do not allow Belsey's findings to be dismissed as Hartley appears to suggest that they should be.

There may still be psychology departments in which blind marking has yet to be adopted or where staff have not yet persuaded their colleagues in other departments to follow suit. In these instances, staff may be interested to set up their own investigations to contribute some data relevant to questions of bias. As I have suggested elsewhere (Bradley, 1993a), such departments might arrange for one marker to mark blind while another marks named scripts. The data could be explored for patterns of sex bias and other biases. Several such departments could join forces to increase opportunities for sub-group analyses, increase generalizability and protect anonymity. I should be glad to put any volunteers in touch with each other and to encourage further appropriate research with a view to promoting examination procedures that are fair.

References

Belsey, C. (1988) Marking by numbers. *AUT Woman, No.15* (autumn).

Bradley, C. (1984) Sex bias in the evaluation of students. *British Journal of Social Psychology*, 23, 147-153.

Bradley, C. (1993) Sex bias in evaluation overlooked? *Assessment and Evaluation in Higher Education*, 16, 1.

Bradley, C. (1993a) Sex bias in evaluations at college and at work: a comment on Archer's review. *The Psychologist: Bulletin of the British Psychological Society*, 6, 56-60.

Hartley, J. (1992) Sex bias, blind marking and assessing students. *Psychology Teaching Review*, 1, 2, 66-75.

Parry-Langdon, N. (1990) "Marking by numbers": evaluation of the marking of final degree examinations in the Faculty of Humanities and Social Studies. Unpublished report for the Faculty and the Deans Committee, University of Wales College of Cardiff.

Manuscript received: 26 November 1992

Final version accepted: 25 March 1993

Address for correspondence: C. Bradley, Department of Psychology, Royal Holloway, University of London, Egham Hill, Egham, Surrey TW20 0EX