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

Sex bias in the evaluation of students

Abstract

Sex differences in academic achievement have often been observed but have usually been attributed to inherent differences between men and women students rather than to differential marking by examiners. The present study indicates that this may not be the case. Sex bias was shown to occur in the evaluation of students' projects. These results cannot be attributed to actual differences in achievement of men and women students. The implications of these findings for marking in general are discussed and recommendations are made for examination procedures which would help to eliminate such bias.

Introduction

There is now considerable evidence to suggest that sex bias occurs in many circumstances of evaluation such as employee selection and promotion. Most of the studies of sex bias in evaluation have examined the hypothesis that men are evaluated more favourably than women when both sexes have identical qualifications or performance. Although many studies have demonstrated such pro-male evaluation bias, some have found no sex bias and others have demonstrated pro-female evaluation bias. Nieva and Gutek (1980) reviewed the literature concerned with sex bias in a wide variety of conditions where men and women are evaluated and suggested that the degree and pattern of bias depends on three factors:

Level of Inference. Bias, including sex bias, tends to operate where there is ambiguity concerning evaluation criteria. The more task-related information provided about the individual to be evaluated and the greater the clarity of the evaluation criteria, the less likely is the operation of 'actuarial' prejudice, (e.g. Pheterson *et al.*, 1971; Teborg and Ilgen, 1975; Hall and Hall, 1976).

Sex role incongruency. Sex bias tends to occur when the tasks undertaken are deemed to be more appropriate for one sex than for the other. Pro-male bias is common in evaluation of performance of traditionally masculine tasks (e.g. Mischel, 1974; Dipboye *et al.*, 1975; Teborg and Ilgen, 1975) while pro-female bias may occur in evaluation of performance of traditionally feminine tasks (e.g. Mischel, 1974; Feather, 1975; Feather and Simon, 1975; Cash *et al.*, 1977).

Level of Performance. The operation of sex bias appears to be affected by the level of qualification or performance involved. Women tend to be evaluated less favourably than men when both men and women are highly qualified or perform well. When qualifications of both sexes are low or their performance is poor, women tend to be evaluated more favourably than men (Deaux and Taynor, 1973; Feather and Simon, 1975).

The evaluation of students' work is supposed to be objective and merit based; however, the evaluation criteria for assessing students' written work are highly ambiguous and the

marking process has long been known to be unreliable (e.g. Hartog and Rhodes, 1935; Dale, 1959; Robbins, 1963; Cox, 1967). Thus a high level of inference is required in evaluating students' written work and sex bias would be expected to occur under such conditions. Because there is a tendency for successful academic performance to be seen as consistent with masculine sex role stereotypes, it was hypothesised that pro-male bias would occur where more able students were concerned. Where the level of performance of both men and women students was low it was hypothesised that pro-female bias would occur. Across the entire range of abilities, therefore, it was expected that the work of men students would be marked more extremely than the work of women students.

Sex differences in the distribution of examination marks are common but the differences are usually attributed to inherent differences in the abilities of men and women students (e.g. Dale, 1959; Murphy, 1982). The present study was designed to exclude the possibility that actual differences in the performance of men and women students would account for the sex differences observed. The marks investigated were those awarded for student projects by two independent markers. The first marker was the project supervisor who knew the student well and was also involved with the planning of the student's work. The second marker had had considerably less contact with the student and was less familiar with the area of study. The second marker would, however, be aware of the student's name and hence the sex of the student. The second marker's more limited knowledge of the student as an individual together with less detailed knowledge of the project area would serve to increase the level of inference required by the second marker compared with that of the first marker. The literature described above would suggest that bias arising from sex-role expectations would be more apparent in the second marker who had less information available on which to base an evaluation.

Method

Project marks from first and second markers in four separate university departments were investigated. All the markers were aware of the names and hence the sex of the student whose projects they were assessing. Markers allocated marks independently before they discussed the work. Analysis was restricted to those marks awarded to projects for which, initially, there was disagreement over the class to be awarded. Percentage disagreement was high (60%, 43%, 42% and 41% for departments 1 to 4) supporting the expectation of ambiguous evaluation criteria. Disagreements across classes were examined rather than precise quantitative differences between first and second markers' marks since the marking scale was not linear. The difference between a mark of 52 and one of 58 is quantitatively greater than the difference between 59 and 60. However, in many universities, 52 and 58 are both lower second class marks while 59 and 60 span the boundary between the lower and upper second classes. Hence the difference of one mark at the class boundary is qualitatively greater than the difference of six between two marks which both fall within the same degree class. It is the qualitative difference that is ultimately of significance to students and examiners: whether a student receives an upper or a lower second class degree is of greater importance than whether they receiver a high or a low mark within either of these classes.

Data were obtained from a fifth department at a Polytechnic where records of marks given by first and second markers for student projects were available for the previous four years. The marking procedures followed by this department were similar to those of the four other departments. There was one important difference: the second markers were unaware of the name and sex of the students. In this department, where the second marker could not be influenced by sex bias, the possibility remained for sex bias in the marking by the first marker although, for reasons outlined above, sex bias in supervisors' marking was not anticipated. The data from department 5 were used to investigate the less likely possibility that the first marker might also be demonstrating sex bias. The sex of the marker was not considered as a variable in the present analyses. Too few women examiners were involved to allow analysis by sex of marker.

Results

The data from each of the departments 1 to 4 were pooled for preliminary analysis of sex bias using Fisher's exact test for 2 x 2 contingency tables. The number of occasions on which the second marker marked more extremely relative to the first marker, i.e. away from the middle of the lower second class, was compared with the number of occasions on which the second marker marked towards the mid-point of the lower second class for men students' projects and for women students' projects. The frequencies, presented in Table 1 suggest that, as predicted, second markers marked men students' projects more extremely while they showed a central tendency when marking women students' projects. Fisher's test showed the results to be statistically significant ($p<0.02$).

Table 1. Pooled data from departments 1, 2, 3 and 4 showing frequencies of second marker marking towards the extremes or towards the centre relative to the first marker for the projects of men and women students over which there were initial disagreements across any one of the class boundaries about the class of the mark.

		Second marker	
		Marked towards centre	Marked towards extremes
Sex of student			
	Men	9	15
	Women	33	16
Fisher's exact test $p<0.015$			

Some statisticians might be concerned that the pooling of data from different sources, in this case from different departments, was inappropriate and misleading, artificially reducing the probability level computed. A more conservative analysis of the data was, therefore, carried out. The probabilities of the observed frequencies for each of the four separate departments for men and women separately were computed using the Binomial Test. The probabilities obtained were converted to z scores, weighted to allow for differences in sample size and combined according to the method recommended by Rosenthal (1978).

Table 2 summarises the data from each of the first four departments showing, for the cases where there were disagreements across classes, the probability of the second marker marking in the predicted direction, i.e. towards the extremes for men and towards the centre for women. When the data for men and women from all four departments were combined by Rosenthal's method, an overall probability level of $p<0.03$ was obtained. Thus, using the more conservative test of significance, the data still provided support for the hypothesis that sex bias was operating in the marking of these student projects.

Table 2. Summary of data from departments 1 to 4 including z scores and probabilities for the marking patterns in individual departments and for men and women students' projects separately.

Dept	Sex of student	Total number of students	Disagreements across classes: second marker marked		z	p
			Towards centre	Towards extremes		
1	Men	20	4	10	1.34	0.09
	Women	35	12	7	0.92	0.18
2	Men	7	1	3	0.49	0.32
	Women	16	3	3	-0.40	0.66
3	Men	12	3	1	-1.53	0.94
	Women	33	11	4	1.56	0.06
4	Men	8	1	1	-0.68	0.75
	Women	19	7	2	1.34	0.09

Note. Combined z score (combined according to the method recommended by Rosenthal, 1978) = 1.884; $p<0.03$

Predictions for department 5, where the second marker was not aware of the name or sex of the student, were the reverse of the predictions for the other departments. The data from department 5 where second markers marked 'blind' were collected over a period of four years and have been pooled for the purpose of the present analysis. These data are described in Table 3. When Fisher's exact test was used to test the hypothesis that first markers showed a central tendency effect in marking the projects of women students and marked the men's projects more extremely, there was found to be no difference between the markers in this respect ($p<0.33$). Thus, in department 5, there was no evidence that first markers demonstrated sex bias in their marking compared with the second markers.

Table 3. Data from department 5 showing frequencies of first marker marking towards the centre or towards the extremes relative to the second marker for the projects of men and women students over which there was initial disagreement about the class of mark to be awarded.

Sex of student	Total number of students	Disagreements across classes	
		First marker marked towards centre	First marker marked towards extremes
Men	80	19	25
Women	75	14	25
Fisher's test p=0.3262, n.s.			

Note. Eighty three cases of initial disagreement across classes are presented here. Three further cases, where both markers were equally extreme, were omitted from this analysis.

It was apparent, however, that there were consistent differences in the marking patterns of first and second markers in department 5. The second markers marked lower than the first markers for both men and women students on 63 occasions while they marked higher than the first markers on only 23 occasions (binomial test; $p<0.0002$, two-tailed test).

The data from departments 1, 2, 3 and 4 were examined in more detail to estimate the likely effects of the sex bias observed. In department 3 the initial independent marks of the first and second markers went forward as two separate marks to the final examiners' meeting. Thus any bias shown by the second markers in department 3 would be directly reflected in the profiles of marks which determined final degree class attained. In departments 1, 2 and 4, however, the two markers brought their initial independent marks to discussion with a view to agreeing a final mark. If no agreement could be reached, advice was sought from an internal moderator or the external examiner.

In departments 1, 2 and 4 it was important to determine the relative influence of the two markers in arriving at the agreed mark. If it were the case that the second marker deferred to the first marker where there was initial disagreement, the second marker's bias would be eradicated. For those cases were agreed marks were available, the initial marks were examined to determine which marker's mark more closely approximated to the final agreed mark. The frequency data presented in Table 4 show no significant difference between the number of occasions on which agreed marks corresponded more closely to the first markers' initial marks rather than to those of the second markers (binomial test $p<0.226$, n.s.). It was not the case, therefore, that any bias shown by the second marker would have been eradicated by the greater influence of the first marker; both markers were equally influential in determining the agreed mark.

Table 4. Number of occasions where the agreed mark more closely approximated the initial marks of first or second markers for cases where marking was or was not in the direction of bias predicted.

	Agreed mark more closely approximated the initial mark of:			
	Second marker	Equidistant	First marker	Totals
Second marker marked in the direction of bias	13[a]	4	17[a]	34
Second marker did not mark in the direction of bias	6[a]	3	8[a]	17
Totals	19[b]	7	25[b]	51

[a] Fisher's test on cell frequencies p= 0.6184, n.s.
[b] Binomial test on column totals p= 0.2257.

Note. Data were from departments 1, 2 and 4 where agreed marks were available. (Agreement could not be reached on three of the projects which were sent to the external examiner).

Discussion

The data indicated that the second markers showed sex bias relative to the first markers when both markers were aware of the sex of the student. When the second marker marked unnamed projects (department 5) no sex bias was apparent; it was not the case that the first marker showed sex bias relative to the second marker. These results suggested that familiarity with the student gained during the period of supervision and/ or the first marker's greater knowledge of the project area, served to eliminate any tendency for sex-role expectations to influence the evaluation of the students' work.

There are several possible explanations for the tendency for second markers to mark down in department 5. The second marker may be more willing to risk downgrading a good project than to risk his or her reputation by upgrading a bad project; the first marker may feel the need to justify his or her research areas and supervisory skills by marking students more generously; the first marker may be more aware of, or more sympathetic to, the difficulties experienced by the student carrying out the project; the second marker may overestimate the extent of the supervisor's assistance. It was not possible to determine the relative contributions of these possible explanations on the basis of the present data. The results from departments 1, 2, 3 and 4, however, showed that the effects of sex bias were sufficient to reverse the tendency for second markers to mark down for more competent men students and for less competent women students.

The implications of sex bias in examining are disturbing. If sex bias occurs not only in the assessment of projects but also in the other examinations which contribute to the final degree classification, individual students may be seriously affected. Women students may be receiving lower second class degrees while men students with comparable abilities are awarded upper second degree classes. Weaker men students

may be penalised more severely than their female counterparts. Before the importance of the effect of sex bias in the examining of any one academic department can be estimated, information is needed about the opportunities for sex bias to operate in the marking of other papers. If all papers were prepared and marked in the manner of the projects considered here, then sex bias would be equally likely to occur. If, however, these additional papers were marked by examiners who had had little contact with the students then there would be even greater opportunity for bias to accumulate. The risk of sex bias may, therefore, be greater in departments with large numbers of students and relatively little staff-student contact. Systematic investigation is required of the effects of student numbers and staff-student contact together with the effects of other factors, including subject of degree course and the ratio of men to women students which may mediate the occurrence of sex bias. The sex of the examiner is probably of less importance in determining the occurrence of sex bias than the traditionality of the examiner as both men and women examiners are exposed to the same cultural stereotypes and expectations of sex-role appropriate behaviour. The extent to which the examiners embrace traditional attitudes towards sex-roles may well be an important mediating factor which requires investigation.

Sex bias in examining could be eliminated if scripts to be assessed were numbered rather than named. Although 'blind' marking is not always possible when first markers have also supervised the students' projects, the present data showed that the first markers were less susceptible to bias than the second markers. Second markers could be given numbered, unnamed projects. Where the marking of written examinations is concerned, the use of candidate numbers would effectively reduce the possibilities of sex bias. Unless scripts are typed, the occasional candidate with eccentric handwriting would be identified but the majority of students in large departments, where sex bias would be expected to be most pronounced, would not be recognised by the markers. Every examiner must be aware of the possibilities of bias and some universities and polytechnics have for years been using candidate numbers as a means of reducing 'halo' effects which may be apparent when examiners are influenced by their expectations of individual students.

While 'blind' marking may eliminate sex bias and 'halo' effects in the evaluation of students' work, the data from department 5 suggest that there is a possibility that examiners marking numbered papers will assign lower marks than would be assigned to named papers. This possibility should be considered by departments contemplating the use of 'blind' marking.

In demonstrating not just isolated, individual bias effects, but a clearly recognisable pattern of sex bias, the present investigation offers compelling reasons for identifying examination procedures where sex bias may operate and for adjusting those procedures to improve the validity of student assessments.

Acknowledgements

I wish to thank Adrian Simpson and David Shapiro for their statistical advice and Peggy Bradley for valuable comments on the manuscript. I also wish to thank the members of staff in the five departments studied who allowed access to their examination marks and to thank particularly the examiner in each of those departments who sought departmental approval on my behalf and provided the data required.

References

Cash, TF, Gillen, B and Burns, DS (1977). Sexism and beautyism in personnel consultant decision making. *Journal of Applied Psychology*, **62**, 301-310.

Cox, R (1967). Examinations and higher education: A survey of the literature. *Universities Quarterly*, **21**, 292-340.

Dale, RR (1959). University standards. *Universities Quarterly*, **13**, 186-195.

Deaux, K and Taynor, J (1973). Evaluation of male and female ability: Bias works two ways. *Psychology Reports*, **31**, 20-31

Dipboye, RL, Fromkin, HL and Wilback, K (1975). Relative importance of applicant sex, attractiveness and scholastic standing in evaluations of job applicant resumes. *Journal of Applied Psychology*, **60**, 39-43.

Feather, NT, (1975). Positive and negative reactions to male and female success and failure in relation to the perceived status and sex-typed appropriateness of occupations. *Journal of Personality and Social Psychology*, **31**, 536-548.

Feather, NT and Simon, JG (1975). Reactions to male and female success and failure in sex-linked occupations: impressions of personality, causal attributions, and perceived likelihood of different consequences. *Journal of Personality and Social Psychology*, **31**, 20-31.

Hall, FS and Hall, DT (1976). Effects of job incumbents' race and sex on evaluations of managerial performance. *Academy of Management Journal*, **19**, 476-481

Hartog, Sir Phillip and Rhodes, EC (1935). *An examination of examinations.* London: Macmillan Press.

Mischel, HN (1974). Sex bias in the evaluation of professional achievements. *Journal of Educational Psychology*, **2**, 157-166.

Murphy, RJL (1982). Sex differences in objective test performance. *British Journal of Educational Psychology*, **52**, 213-219.

Nieva, VF and Gutek, BA (1980). Sex effects on evaluation. *Academy of Management Review*, **5**, 267-276

Pheterson, GT, Kiesler, SB and Goldberg, PA (1971). Evaluation of the performance of women as a function of their sex, achievement and personal history. *Journal of Personality and Social Psychology*, **19**, 114-118.

Robbins, LC (1963). *Report of the committee on Higher Education*. Cmnd 2154 (Robbins Report). London: HMSO.

Rosenthal, R (1978). Combining results of independent studies. *Psychological Bulletin*, **85**, 185-193.

Terborg, JR and Ilgen, DR (1975). A theoretical approach to sex discrimination in traditionally masculine occupations. *Organisational Behaviour and Human Performance*, **13**, 352-376.