Does the gender of examiners influence their marking?

A paper to accompany a poster presented at the conference *Learning Communities and Assessment Cultures: Connecting Research with Practice*. The conference was jointly organised by the EARLI Special Interest Group on Assessment and Evaluation and the University of Northumbria.

Jackie Greatorex and John F. Bell

University of Cambridge Local Examinations Syndicate

Disclaimer
The opinions expressed in this paper are those of the author and are not to be taken as the opinions of the University of Cambridge Local Examinations Syndicate (UCLES) or any of its subsidiaries.

Note
This research is based on work undertaken by the University of Cambridge Local Examinations Syndicate for Oxford Cambridge and RSA Examinations (OCR).

Contact details
Jackie Greatorex
Research and Evaluation Division, University of Cambridge Local Examinations Syndicate, 1 Hills Road, Cambridge, CB1 2EU.
☎ 01223 553835
FAX: 01223 552700
💻 greatorex.j@ucles.org.uk www.ucles-red.cam.ac.uk

Does the gender of examiners influence their marking?

Abstract

For many years there have been concerns about whether sex bias exists in various assessments. A literature review reveals that little if any sex bias has been found in UK national assessments at General Certificate Secondary Education (GCSE) mostly taken by 16 year olds and General Certificate of Education mostly taken by 18 year olds. Oxford Cambridge and RSA examiners, for three case study subjects, completed the Bem Sex-Role Inventory (BSRI). This self-report inventory measured the extent to which the respondents are sex-typed i.e. to what extent they have socially desirable, stereotypically masculine and feminine personality traits. The responses were used to investigate whether there was any:-

- sex bias in the marking;
- relationship between the masculinity and femininity of the examiners and their marking of male and female examinees.

It was found that there was one question in the English examination which was biased by 0.5 of a mark in the favour of girls. There was another significant result in English, the more senior the examiner the more generous the marking. These concords with results in other areas like testing English as a foreign language. It is concluded that sex bias should be monitored but is unlikely to be found.

Introduction

Sex bias

Over many years studies have focused on whether the whole test or individual questions and/or associated mark schemes are 'biased'. The concern is that particular groups, whether they are gender, ethnic or other types of groups gain lower marks than other groups. For example, a major concern for many years was that men gained higher marks than women did on the Scholastic Aptitude Test (SAT) in America (see 'The FairTest Examiner Volumes 1 to 5 ‘1987 to 1991'). It is a matter of professional judgement to determine when a disparity in marks is a bias.

Other studies have focused upon another form of sex bias or gender bias. Gipps (1994) explains that bias can occur when the score given by an examiner is consciously or unconsciously affected by factors other than the candidates’ achievement e.g. sex, ethnic origin, school, handwriting. Alternatively markers award marks to answers, which illustrate skills, knowledge and/or values irrelevant to the test but which are valued by the markers. There is some evidence of this kind of sex bias, for example, O' Neill (1985) found that in teachers' assessment of student work markers devalued the performance of their own sex. Additionally, it has been found that examiner behaviour varies with different groups, such as professional background, subject specialism and gender (Hamp-Lyons, 1990; Vann Lorenz and Meyer, 1991). This is presumably due to each group having a unique frame of reference.

Sometimes this issue of sex bias is investigated through blind marking studies. Spear (1984) found that teachers awarded higher ratings to work that was believed to be completed by boys. Newstead and Dennis (1990) compared the marks awarded to psychology graduates in two UK institutes of Higher Education. One of the institutes used blind marking. Although Newstead and Dennis found blind marking did not affect the achievement or marks awarded to males and females Bradley (1993) re-analysed the data and found a pro-male bias. Swim et al. (1989) found that both male and female academics rated an article as better when it was believed to be by John T McKay than when the author was believed to be Joan T McKay. Gipps and Murphy (1994) found that there were interaction effects between the sex of the marker and the sex of the assessee in Craft, Design and Technology (CDT), that is markers awarded more marks to assessees of their own sex. However, Gipps and Murphy (1994) argued that this was really a subject bias – the female markers were Home Economics teachers by training and the male teachers were trained to be CDT teachers.

In England there are three Awarding Bodies (Assessment and Qualifications Alliance (AQA), EdExcel and Oxford, Cambridge and RSA Examinations (OCR)) that administrate general and occupationally-related qualifications. The general qualifications include the General Certificate of Secondary Education (GCSE), which is a national assessment normally taken in a series of subjects by sixteen year olds. There is also the General Certificate of Education Advanced Level (GCE A Level), which is a national assessment, normally taken in three or so subjects by eighteen year olds.

Baird (1996) investigated sex bias in marking in Chemistry and English Literature A level using a blind marking approach. She found that marks were not affected by the gender of the examinee or the style of handwriting in either subject. In the case of 'live' GCE and GCSE examinations blind marking would be a considerable logistical challenge.

Of course blind marking is not truly blind as the assessed work includes cues to the sex of the assessee, for instance the style of handwriting. Girls are perceived to have neater handwriting and Woods (1991) reported that neatness of presentation and clear handwriting affect marks in an upwards direction. So even in a blind marking scenario the sex of the examinee might be inferred and stereotypes can come into play. The Scottish Examining Board (1992) investigated marker practices in English and History. In this study examiners marked scripts which varied in the achievement of the centre, handwriting, gender and ethnic origin of the candidate. The only significant effect found was that the typed scripts gained lower marks than the hand written scripts.

The first issue researched here is whether male and female examiners respond differently to the performance and answers of candidates of different sexes. It is beyond the scope of this paper to consider the causes of any differences. For example, there is the issue of face validity or biased question context, which is a complex issue. For further information on this see Pollitt and Ahmed (2000).

Androgyny

Whilst there is a great deal of literature about sex bias (using biological sex) there seems to be comparatively little work linking the issue of androgyny to sex and /or gender bias in educational assessment. Indeed Morgan (2002) argues that generally within sociological work masculinity and feminity appear to be an assumed extension of the sex of those studied, and there is little work which attempts to separate masculinity and feminity from sex.

Constantinople (1973) revolutionised the way that masculinity and femininity were conceptualised when she argued that they were independent constructs and not opposite ends of a unidimensional continuum. From this argument followed psychological androgyny theory, with the principle that individuals could be both masculine and feminine (Ballard-Reisch and Elton, 1992). However, once people in general are androgynous masculinity, femininity and androgyny cease to be meaningful and the theory becomes redundant (Ballard-Reisch and Elton, 1992). Eichler (1980) criticises androgyny by saying that it assumes that there are some standards of what constitutes masculinity and feminity against which deviations are recorded, therefore a theory which is associated with promoting equality is actually problematising people who deviate from having the personality traits stereotypically associated with their sex.

The second issue researched here is whether the masculine and/or feminine examiners respond differently to the performance and answers of candidates of different sexes.

Bem (1974, 1979) argued that an androgynous gender (sex-role) orientation enables people to be flexible and to respond appropriately in a variety of situations. Additionally one's psychological wellbeing is maximised if you are androgynous (Whitley, 1984). In contrast the congruence model asserts that the healthiest, most competent orientation for an individual is one corresponding to his/her sex.

In this paper both the biological sex and the gender of examiners (of GCSE English, History and Food Technologies) will be considered. The gender (masculinity and/or feminity) of the examiners is described in and measured by the Bem Sex-Role Inventory.

Bem Sex-Role Inventory

The Bem Sex-Role Inventory (BSRI) measures the masculinity and femininity of individuals. Respondents are categorised as:-

- masculine (high on the masculine scale and low of the female scale);
- feminine (low on the masculine scale and high on the feminine scale);
- androgynous (high on the masculine and feminine scales);
- undifferentiated (low on the masculine and feminine scales).

Each scale constitutes a list of traits to which the participants respond with an indication of how well each trait describes them. Some of the traits are not part of the masculine and feminine scales- they are 'fillers'.

The Bem Sex-Role Inventory was developed in the 1970s. Therefore some might reasonably question whether studies based upon its use are still valid. Harris (1994) found that within limits the BSRI is a valid measure of masculinity and femininity in American culture. Harris (1994) also found that the BSRI was less valid for Hispanic-American and African-American people than it was for Anglo-American people. Auster and Ohm (2000) reviewed the literature about androgyny and societal changes. They concluded that whilst many societal changes have taken place in terms of gender roles, e.g. married women's increased participation in the work force, men and women still believe that American society in general still desires men to be masculine and women to be feminine as measured by the BSRI. Ballard-Reisch and Elton (1992) found along with other authors that some 'neutral' characteristics in the BSRI (the fillers) have since been considered to be masculine and feminine. They found that the scales in the BSRI were reliable but questioned whether they measured masculinity and femininity. But Auster and Ohm (2000) say that *The findings of our study bring the validity of masculine and feminine dimensions of the BSRI into question if we evaluate the masculine and feminine traits with the criteria used in the original development of the instrument. On the other hand, the striking patterns of the desirability ratings arranged in rank order might cause one to be less critical of the BSRI because the ratings of the traits seemed so traditionally gender typed…..The respondents perceptions of most men and most women were quite gender typed* (Auster and Ohm, 2000, 525). This suggests that the masculinity scales and femininity scales still have some validity in modern American society.

Murray (1976) found that ratings of psychological health for women varied as a function of the task in which the woman was portrayed and the rater's sex-role identity. Downing (1978) found that masculine, feminine and androgynous individuals described the psychologically health of males and females differently from one another. Downing's study along with other studies e.g. Spence et al. (1975) illustrates that self-perceptions tend to be less stereotypic than perceptions of an imagined male or female. Additionally Delia (1972) found that the amount of information an individual knows about the person is inversely proportional to the amount of stereotyping the individual uses in forming perceptions of that person. Bradley (1984) argued her findings showed that greater knowledge of the student reduced the incidence of sex bias in university assessments. This would lead to the possibility of there being more sex bias in externally set and assessed components of qualifications like GCSEs than the coursework part of GCSEs. However there are other biases which might be at work in the marking of coursework.

Another limitation of the Bem Sex-Role Inventory is that it measures a series of gendered personality traits (Blanchard-Field, Suhrer-Roussel and Hertzog, 1994). But personality traits are not the whole story, there may be other components which constitute gender role, for example, attitudes, stereotypes, behaviours, social relationships, abilities and interests (Ashmore, 1990). Additionally it can be argued that negative aspects of undertaking research about masculinity and feminity serves to perpetuate the stereotypes which the work might be aimed at eradicating (Morgan, 2002).

This study

Despite its limitations the Bem Sex-Role Inventory was chosen as a useful tool because Auster and Ohm (2000) argue that it still has some validity. Additionally other possible measures of masculinity and femininity (amongst other traits) such as the Minnesota Multiphasic Personality Inventory are for diagnosing personality disorders. Such an instrument would be wholly unsuitable for a research study about employees completed by an employer. A limitation of an employer sending a questionnaire to it's employees is that they are unlikely to respond in socially undesirable ways, for example, people might be more likely to respond in a gender stereotyped manner if they believe that this is more socially desirable from their employers' perspective. It should be noted that respondents were told that their individual responses were confidential to the Research and Evaluation Department (RED) of UCLES and that OCR would not have their individual data. Of course, examiners might not trust the assurances given by RED.

The literature therefore indicates that there are some relationships between individuals' self perception (gender) as measured by the Bem Sex-Role Inventory and their reactions to males and females or imagined males and females. But it was also found that the more information an individual has about someone else the less stereotypes come into play. The aim of this study is to investigate any relationship, which might or might not exist between the self-perception of masculinity and femininity (gender) of examiners and their marking of male and female examinees.

In this study candidates' marks and examiners were used from three GCSEs:-

- English 1500 paper 3 (Non-Fiction and Media Texts) Higher Tier paper, comprising 4 questions of which 1 and 2 were compulsory but candidates could choose between questions 3 and 4, each question was worth 20 marks;
- History 1605 paper 2 (Medicine Through Time), including 7 compulsory short essay questions worth between 6 and 12 marks each. The questions were based around a series of 8 'sources', i.e., a map, 4 short extracts from books and pictures of a Roman Coin, a Greek carving and a Roman carving;
- Design and Technology: Food Technologies (Food) 1460 paper 1 a Foundation Tier paper constituting 5 compulsory short answer questions (or sub questions) worth no more than 4 marks each.

Each subject in higher education has it's own skills, knowledge domain and culture (Becher, 1989). Indeed these different areas might lead to different types of assessment being used (Woolf, 2000). These principles also apply to the GCSE subjects English, History and Food. They are all quite different in terms of knowledge and skills and this is reflected in the different types of question papers and mark schemes. It is also noticeable in the stereotypical numbers of examiners in each subject team Table 1 and in examining as a whole, where the senior examiners tend to be male even in predominantly female subjects (Moody, 1999). An exception to this general trend is Food where all the examiners are female but there are some male candidates (see Table 1).

This paper is a report of one part of a programme of research about the gender of examiners inspired by Moody (1999). The other part of the research is about the careers and seniority of the roles held by examiners of different sexes and sex-role orientations which will be reported in Greatorex and Bell (2002).

Method

Sample

There were four samples of examiners and associated samples of candidates' work marked by these examiners. The first three samples constituted the unit (examination paper) level marks awarded by examiners to the candidates in three GCSE subjects listed above. The fourth was a sample of the item (question) level marks awarded by some examiners to candidates in an English unit.

Unit level data

The Bem Sex-Role Inventory (detailed below) was mailed to all examiners who had marked the GCSEs listed above in the summer 2001 session.

Table 1 Distribution of examiners who marked each paper by sex

Subject	Female	Male	Unknown	Total
English	76	60	6	147
Food	82	0	0	82
History	21	35	1	57

English was used as there was a fairly even number of male and female examiners, History as there were more male than female examiners and Food as there were no male examiners but candidates of both sexes.

Table 2 Distribution of examiners who returned the questionnaire by sex

Subject	Female	Male	Unknown	Total
English	48	56	0	104
Food	53	0	0	53
History	10	25	0	35

The response rate was 70% for English, 65% for Food and 61% for History, which are all acceptable.

Item Level Data

The item level data was collected in the autumn term of 2001. When the item level data was collected from the English scripts many of the scripts were unavailable to be used for research purposes as they were being used for operational purposes. So the choice of scripts that could be used was significantly affected, as operational purposes are prioritised over research purposes. For example, there were some examiners whose entire allocation was being used for operational purposes and others for whom the majority of the scripts were not available.

133 examiners of 142 examiners' marking were sampled. The 7th script from each examiners' marking was taken from scripts which are stored in examiner, centre and then candidate number order. The data at the question (item) level was keyed into a database. If there were a small number of scripts then 7th was replaced by 4th. This was repeated 20 times for each examiner. If fewer than 20 scripts for an examiner were available then they were excluded from the sample. The lack of scripts available meant that the sample of examiners' scripts was often taken from one centre. However there were some examiners who had an allocation of only one centre. In other studies where this method of sampling was checked for systematic

biases it was concluded that systematic biases were not introduced. But the sample of examiners in Table 1 was not necessarily randomly selected from all the examiners who marked that examination. The examiners whose marking was available might be the marking most similar to that which centres expected. Awarding Bodies know what outcomes the centres expect for each candidate as they give forecast grades and request a small proportion of remarks.

Table 3 Summary statistics for English

Statistic	Entry	With questionnaire data	With questionnaire and item data
No. of examiners	142	104	83
No. of centres	638	566	87
No. of candidates	42,396	31,738	3,958
Mean mark	17.3	17.3	17.5
Standard deviation of marks awarded	4.1	4.1	4.2

Based upon the mean marks awarded and the standard deviation of these marks in Table 3 both samples of English examiners were representative of the examiners who marked English in the summer 2001 session.

Procedure

Bem Sex-Role Inventory

The Bem Sex-Role Inventory was mailed to all examiners listed above. These examiners were told that the individual results of the Bem Sex-Role Inventory would remain confidential to the Research and Evaluation Division (RED) of University of Cambridge Local Examinations Syndicate (UCLES) and that they would be reported as a group. The examiners received individual letters informing them of their results from the BSRI. A minority of examiners wrote to the authors to ask for clarification about the research and these examiners received individual letters answering their queries.

Unit level data

Awarding Bodies store a variety of operational data including the marks achieved at the unit level by each candidate in each qualification and the candidates' sex. These data were matched with the examiners for each paper to undertake the analysis described below.

Item level

Awarding Bodies do not store operational information about the item level marks gained by GCSE candidates as a matter of course (for instance scores for multiple choice questions might be stored). These data were collected from the scripts as detailed above.

Analysis

A multilevel model was used to explore the scores awarded by the examiners at the unit and item level in relation to:-

- the sex of the examiner and the examinees;
- the gender of the examiner and the sex of the examinees.

Results

Bem Sex-Role Inventory

Table 4 Distribution of Examiners by sex-role group

Subject	Androgynous	Feminine	Masculine	Undifferentiated	Total
English (item)	9	17	15	27	68
English (unit)	16	26	24	38	104
Food	12	14	9	18	53
History	7	1	13	14	35

An analysis of the masculine and feminine scores derived from the BSRI revealed one significant difference (Greatorex and Bell, 2002). The female History examiners had a lower average score of the femininity index than the female examiners in the other two subjects. This explains why the BSRI has classified so few History male and female examiners as feminine.

Item level data

English

Item level data was only available for English. This was analysed using a multilevel model. There were two limitations with the data set. Firstly, although the candidates were offered a choice between questions 3 and 4, almost all candidates attempted question 3 (and no-one in the sample of item level data attempted question 4). This simplifies the modelling process. Secondly, in the sample of scripts in almost all cases all the scripts marked by a given examiner were from the same centre, this means that it was only possible to fit a two level model; examiner/centre and candidate.

Table 5 Item level multilevel model results for English

Fixed Part

	Intercept		Question 2		Question 3		Examiner Sex		Q2*csex	
Model	Est	s.e.	Est	s.e.	Est	s.e.	Est	s.e.	Est	s.e.
I = Null	11.3	0.2	-	-	-	-	-	-	-	-
II = I + questions	11.5	0.2	-0.9	0.1	0.5	0.1	-	-	-	-
III = I + exam_sex	11.6	0.3	-0.9	0.1	-0.5	0.1	-0.2	0.4	-	-
IV = III + q2*candsex	11.5	0.2	-0.6	0.1	0.5	0.1	-	-	-0.5	0.2

Random part

	Examiner/Centre			Candidate		
Model	Est	s.e.	s.d.	Est	s.e.	s.d.
I =	3.2	0.6	1.8	8.5	0.2	3.0
II =	3.2	0.6	1.8	8.2	0.2	2.9
III =	3.2	0.6	1.8	8.2	0.1	2.9
IV =	3.2	0.6	1.8	8.2	0.2	2.9

In Table 5; all parameter estimates are significant. In the item level data set, there was no examiner sex difference. Male candidates tended to obtain 0.5 of a mark less on question 2 *Explain how each writer, through his content and use of language, engages the interest of the reader?* (The question paper was accompanied by two articles about tornadoes which the candidates were instructed to read). Models using the masculinity and femininity scores were also fitted in place of the sex of the examiner and there were no significant results.

Unit level data

English

Table 6 Unit level multilevel model for English

Fixed Part

	Intercept		Candidate Sex		Examiner Sex	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
I = null	17.0	0.1	-	-	-	-
II = I +cand. Sex	17.2	0.1	-0.5	0.0	-	-
III = II + exa. Sex	17.3	0.2	-0.5	0.0	-0.1	0.3
IV = II + masc.	17.2	0.1	-0.5	0.0	-	-
V = II +fem.	17.2	0.1	-0.5	0.0	-	-
VI = II + status	17.0	0.2	-0.5	0.0	-	-

Fixed Part

	Masculinity		Femininity		Status	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
IV = II +masc.	0.1	0.1	-	-	-	-
V = II +fem.	-	-	-0.1	0.1	-	-
VI = II + status	-	-	-	-	1.3	0.4

Random Part

	Examiner			Centre			Candidate		
Model	Est.	s.e.	s.d.	Est.	s.e.	s.d.	Est.	s.e.	s.d.
I = null	0.7	0.3	0.8	5.9	0.4	2.4	10.7	0.9	3.3
II = I +cand. Sex	0.7	0.3	0.8	5.8	0.4	2.4	10.6	0.9	3.3
III = II + exa. Sex	0.7	0.3	0.8	5.8	0.4	2.4	10.6	0.9	3.3
IV = II +masc.	0.7	0.3	0.8	5.8	0.4	2.4	10.6	0.9	3.3
V = II +fem.	0.7	0.3	0.8	5.8	0.4	2.4	10.6	0.9	3.3
VI = II + status	0.6	0.3	0.8	5.7	0.4	2.4	10.6	0.9	3.3

In Table 6; the results for English at the unit level are presented, examiner sex, masculinity and femininity were not significant. This is similar to the results of the item level analysis but is based on a larger sample of 104 examiners (the item level analysis was based on 68 examiners). In this case, the status of the examiner was significant. Team Leaders tended to be relatively more generous than Assistant Examiners. A 95% confidence interval for the examiner level variation is ±1.6 marks and examiner variation accounts for 4% of the total variations.

History

Table 7 Unit level multilevel model for History

Fixed Part

	Intercept		Candidate Sex		Examiner Sex	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
I = null	31.1	0.5	-	-	-	-
II = I +cand. Sex	33.4	0.5	-3.0	0.5	-	-
III = II + exa. Sex	31.1	1.5	-3.0	0.5	1.8	1.1
IV = II + masc.	33.4	0.5	-3.0	0.2	-	-
V = II +fem.	33.4	0.5	-3.0	0.2	-	-
VI = II + status	33.6	0.5	-3.0	0.2	-	-

Fixed Part

	Masculinity		Femininity		Status	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
IV = II +masc.	-0.4	0.5				
V = II +fem.			-0.8	0.5		
VI = II + status					-1.2	1.4

Random Part

	Examiner			Centre			Candidate		
Model	Est.	s.e.	s.d.	Est.	s.e.	s.d.	Est.		s.e.
I = null	4.5	2.2	2.1	33.4	3.2	5.8	98.4	1.1	9.9
II = I +cand. Sex	4.3	2.1	2.1	32.2	3.0	5.7	96.4	1.1	9.8
III = II + exa. Sex	3.7	1.9	1.9	32.2	3.0	5.7	96.4	1.1	9.8
IV = II +masc.	4.3	2.1	2.1	32.2	3.0	5.7	96.4	1.1	9.8
V = II +fem.	3.9	1.9	2.0	32.2	3.1	5.7	96.4	1.1	9.8
VI = II + status	4.1	2.0	2.0	32.2	3.1	5.7	96.4	1.1	9.8

The results from the multilevel modelling of the History data set are presented in Table 7. There were no significant examiner, femininity or masculinity effects. For a parameter estimate to be significant it must be greater than or equal to twice the size of the standard error. The largest parameter estimate was for examiner sex. Although this was 1.8 marks, it was not significant. This means that male examiners gave 1.8 marks more than female examiners but that there was insufficient power in the statistics to say that this was generally applicable (it was not significantly different). The examiner variation is only approximately 3% of the total variation. A 95% confidence interval for the examiner level variations was ± 4.2 marks but it should be stressed that this analysis is based on raw marks and quality control procedures have not been applied (although they were in the live examination). The parameter estimate for centre level variation of 33.4 is quite large, but these analyses do not account for the achievement of candidates when they entered the school, and GCSE results tend to be highly correlated with past performance. Consequently this variation is affected by prior achievement and cannot be used to make assertions about centres.

Food

Table 8 Unit level multilevel model results for Food

Fixed Part

	Intercept		Candidate Sex		Masculinity		Femininity		Status	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.	Est.	s.e.	Est.	s.e.
I = Null model	21.3	0.2	-	-	-	-	-	-	-	-
II = I +cand. Sex	21.5	0.2	-0.5	0.1	-	-	-	-	-	-
III = II + masculinity	21.5	0.2	-0.5	0.1	-0.0	0.2			-	-
IV = II + feminity	21.3	0.2	-0.5	0.1	-	-	0.3	0.2	-	-
V = II + status	21.4	0.2	-0.5	0.1	-	-		-	0.1	0.5

Random Part

	Examiner			Centre			Candidate		
Model	Est.	s.e.	s.d.	Est.	s.e.	s.d.	Est.	s.e.	s.d.
I	0.7	0.3	0.8	5.1	0.5	2.3	30.2	0.4	5.5
II	0.7	0.3	0.8	5.1	0.5	2.3	30.2	0.4	5.5
III	0.7	0.3	0.8	5.1	0.5	2.3	30.2	0.4	5.5
IV	0.7	0.3	0.8	5.1	0.5	2.3	30.2	0.4	5.5
V	0.7	0.3	0.8	5.1	0.5	2.3	30.2	0.4	5.5

In Table 8, the results for the multilevel models for the Food examination are presented. In this case, all the examiners were female (however, approximately one third of the entry was male). This means that the models only include the candidates' sex, BSRI measures of masculinity and femininity and the status of the examiners (and omit examiner sex). From the random part of the model, it can be calculated that an proximate 95% confidence interval for the examiner variation is ± 1.6 marks.

Further models were used to explore the interaction between:-

- examiner's sex and candidate's sex;
- examiner's sex-role orientation and candidate's sex.

But they are not given here as there was no interaction effects and there were no significant differences.

Discussion

The results given above for the item and unit level data were based upon an analysis of the raw marks gained by the candidates before quality procedures like scaling were followed. The results from the analysis should be interpreted with this caveat in mind.

Although significant sex differences were found, care is needed in interpreting them because of the effects of tiering. This suggests that as a general rule the sex and gender of examiners and interactions between candidate's sex and examiner sex does not affect the marks that candidates gain at the unit level. In other words although examining is male dominated this has not resulted in a bias against girls or boys in the marking. Given the literature reviewed above and the quality procedures of Awarding Bodies this a positive but not unexpected finding.

However there were two significant differences found in English:-

1. there was a small sex bias on one item;
2. the status of the examiners was a significant factor, the more senior examiners (Team Leaders) were more generous than the Assistant Examiners.

The first finding is important given that the scripts available were likely to be those where the marking was most similar to that which schools and colleges expected. On the other hand the difference that was made to candidates' marks was very small so the bias on this item is unlikely to have affected candidates' grades. It is possible that these differences in achievement are due to question 2 being more girl friendly. Question 2 was *Explain how each writer, through his content and use of language, engages the interest of the reader?* (The question paper was accompanied by two articles about tornadoes which the candidates were instructed to read). It could be that this question is girl friendly as to answer it the candidate might put themselves in the place of the author and/or different audiences. Given that different questions can be girl/boy friendly and lead to an inequality in achievement at the item level this suggests that boy/girl friendly questions should be avoided unless they are an integral part of the subject and removing them would reduce the validity of the examination. Additionally it can be argued that boy/girl friendly questions in examination papers perpetuate gender stereotypes, and are therefore undesirable. On the other hand Kiwan et al. (1999) found that candidates used schemas (familiar scenarios stored in memory) to answer and understand examination questions, especially in a stressed situation. So arguably showing males and females in nonstereotypical roles makes questions less accessible. The relationship between question context and candidate achievement might be an area for future research.

The finding that Team Leaders are more generous markers than the Assistant Examiners suggests that Team Leaders are more confident about following the principle of giving the candidate the benefit of the doubt than the Assistant Examiners. However it might also be explained in another way. Team Leaders are experienced examiners and the average Team Leader might be more experienced that the average assistant examiner. If this estimate is appropriate then this finding links to results in testing English as a Foreign Language where experienced examiners are sometimes found to be more lenient than inexperienced examiners (Ruth and Murphy, 1988; Weigle, 1998). But there is evidence that the severity and leniency of marking by examiners of the same status in UK examinations might vary. For instance, Newton (1996) found that the marking of one Senior English examiner was consistently harsher than three other Senior Examiners when they remarked photocopied scripts. He argues that this might be because of the time lag between standardisation, the live marking and the remarking of the photocopied scripts. Arguably it could also be because this Senior Examiner tended to have the trait of being harsher than his or her colleagues.

One aspect of the examining which has not been accounted for in this study is tiering. It could be that there are more gendered answers and gender/sex biases might occur in the Lower rather than the Higher Tier. This area might be an area for future investigation. The tiers used in this study were chosen as they had the larger number of examiners.

In conclusion sex and gender bias in marking is something which should be monitored in GCSE marking but it is unlikely to be found to an extent that affects grades. It appears that question papers should continue to be scrutinised for girl/boy friendly questions which should arguably be avoided. It appears that any differences in the severity and leniency of marking are due to the factors other than the examiner's sex and gender and/or the candidates' sex. Indeed the greatest source of variance was the candidates' achievement, which is as it should be.

Acknowledgements

We would like to thank the History, English and Food Technologies examiners for completing the BSRI.

References

Ashmore, R. D. (1990) Sex, gender and the individual. In L.A. Pervin (Ed.) *Handbook of Personality: Theory and Research.* New York: Guildford Press.

Auster, C. J. and Ohm, S. C., Masculinity and Femininity in Contemporary American Society: A Reevaluation Using the Bem Sex-Role Inventory, *Sex-Roles*, 43, 7/8, 499 - 528.

Baird, J. (1996) *What's in a name? Experiments with blind marking in A level Examinations*, A paper presented at the British Psychological Society Conference in London on 17 and 18 December.

Ballard-Reisch, D. and Elton, M. (1992) Gender Orientation and the Bem Sex-Role Inventory: A Psychological Construct Revisited, *Sex-Roles*, 27, 5-6, 291-306.

Becher, R. (1989). *Academic Tribes and Territories.* The Society for Research into Higher Education and Open University Press: Milton Keynes.

Bem, S. (1974) The Measurement of Psychological Androgyny, *Journal of Consulting and Clinical Psychology*, 42, 155-162.

Bem, S. (1979) The Theory and Measurement Androgyny. A reply to Pedhazur-Tetenbaum and Locksley-Colten critiques. *Journal of Personality and Social Psychology*, 37, 1047-1054.

Blanchard Fields, F., Suhrer-Roussel, L. and Hertzog, C. (1994) A confirmatory factor analysis of the Bem Sex-Role Inventory: Old questions, new answers*, Sex-Roles*, 30, 5/6, 423 - 457.

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

Bradley, C. (1993) Sex bias in student assessment overlooked? *Assessment and Evaluation in Higher Education*, 18, 1, 3-8.

Constantinople, A. (1973) Masculinity-femininity: An exception to a famous dictum. *Psychological Bulletin*, 80, 389-407.

Delia, J. (1972) Dialects and the effects of stereotyping on interpersonal attraction and cognitive processes in impression formation. *The Quarterly Journal of Speech*, 58, 285-297.

Downing,N. E. (1978) The Broverman Study Revisited: Implications for Androgyny, *A paper presented at the 1978 Annual Convention of the Southeastern Psychological Association.* Symposium title: Sex-Role stereotyping in psychotherapy.

Eichler, M. (1980) The Double Standard: A Feminist Critique of Feminist Social Science, St Martin's Press, New York.

Fairtest Examiner (1991) The 'Enhanced' ACT: Still biased, Inaccurate, Coachable and Misused, *FairTest Examiner*, Fall 1991, 6 to 7.

Gipps, C. V. (1994) *Beyond Testing Towards a Theory of Educational Assessment*, The Falmer Press: London.

Gipps, C. and Murphy P. (1994) *A Fair Test? Assessment, Achievement and Equity*, Open University Press, Milton Keynes.

Greatorex, J. and Bell, J. F. (2002) *What makes a senior examiner?* A paper to be presented at the British Educational Research Association Conference, University of Exeter, 12-14 September.

Hamp-Lyons (1990) *Second Language Writing: Assessment issues.* In B. Kroll (Ed.), Second Language Writing: Research Insights for the Classroom (pp 127-153 Norwood, NJ: Ablex Publishing Coorporation.

Harris, A. C. (1994) Ethnicity as a Determinant of Sex-Role Identity: A Replication study of item selection for the Bem Sex-Role Inventory, *Sex-Roles*, 31, 3/4, 241 - 273.

Kiwan, D., Pollitt, A. & Ahmed, A. (1999) *The effects of stress on text comprehension and performance in examinations.* British Psychological Society London conference, December.

Moody, J. (1999) *Jobs for the boys? An investigation into the under-representation of women in senior examining positions with OCR*, Unpublished MEd dissertation, University of Bristol.

Morgan, D. H. J. (2002), *Gender – Key Variables*, Available at http://qb.soc.surrey.ac.uk/resources/keyvariables/morgan.htm

Murray, B. (1976) *Androgyny and sex-role stereotypes: women's real and self-perceptions and perceptions of psychological health in others.* Doctoral Dissertation, California School of Professional Psychological, 1976. Dissertation Abstracts International, 37, 1444B. University Microfilms No. 76—79, 645.

Newstead, S. and Dennis, I. (1990) Blind marking and sex bias in student assessment, *Assessment and Evaluation in Higher Education*, 15 (2), 132-139.

Newton, P. E. (1996) The Reliability of Marking of General Certificate of Secondary education Scripts: mathematics and English, *British Educational Research Journal*, 22, 4, 405 - 420.

O'Neill, G (1985) Self, teacher and faculty assessments of student teaching performance: a second scenario. *The Alberta Journal of Educational Research*, 31, 2, 88-98.

Pollitt, A. & Ahmed, A. (2000) *Comprehension Failures in Educational Assessment,* A paper presented at the European Conference on Educational Research, University of Edinburgh, 20-23 September..

Ruth, L. & Murphy, S., (1988) *Designing writing tasks for the assessment of writing.* Norwood, NJ: Ablex Publishing Corp.

Scottish Examining Board (1992) *Investigation into the effects of the characteristics of candidates and presenting centres on possible marker bias.* Scottish Examination Board internal report.

Spear, M. G. (1984) Sex bias in science teachers' ratings of work and pupil characteristics, *European Journal of Science Education*, 6, 4, 369-377.

Spence, J., Hemrich, R. and Strapp, J. (1968), Ratings of self and peers on sex-role attributes and the relation of self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 32, 29-39.

Swim, J., Borgida, E. and Maruyama, G. (1989) Joan McKay versus John McKay: do gender stereotypes bias evaluations? *Psychological Bulletin*, 105, 3, 409-429.

Vann, R.J. Lorenz, F. O. and Meyer, D. M. (1991) *Error Gravity: Faculty Response to Errors in the Written Discourse of Nonnative Speakers of English.* In L. Hamp-Lyons (Ed.). Assessing Second Language Writing in Academic Contexts (pp.181-195). Norwood, N.J: Ablex Publishing Corporation.

Weigle, S. (1998) Using FACETS to model rater training effects, *Language Testing*, 15 (2) 263-287.

Whitley, B. E. Jr (1984) Sex-Role and psychological well-being: Two meta-analysis. *Journal of Sex-Roles*, 12, 207 - 221.

Woods, R. (1991) *Assessment and Testing. A Survey of Research*, Cambridge University Press: Cambridge.

Woolf, H. (2001) *Making a Mark: What do we do when we grade assignments?* A paper presented at British Educational Research Association Conference, 7-9 September , University of Cardiff.

Appendix B

Notes about how the results for the multilevel model for History is interpreted

The results from the multilevel modelling of the History data set are presented in Table 9. There were no significant examiner, femininity or masculinity effects. For a parameter estimate to be significant it must be greater than or equal to twice the size of the standard error. For masculinity the parameter estimate of -0.4 is less than 2x 0.5 (the standard error), for feminity the parameter estimate is -0.8 which is less than 2x 0.5 (the standard error) and for status -1.2 (parameter estimate) is less than twice the standard error (2x1.4). All these figures are taken from the fixed part.

The largest parameter estimate was for examiner status. Although this was 2 marks, it was not significant (see above). This means that that team leaders gave 2 marks more than assistant examiners but that there was insufficient power in the statistics to say that this was generally applicable (it was not significant different).

The examiner variation is only approximately 3% of the total variation. The variation for examiners was 4.5 (a parameter estimate in the random part), the total variation was 4.5 + 33.4 +98.4 (parameter estimates in the random part) = 136.3, 4.5 is 3% of 136.3. 3% is very low, and so the candidates' grades are very unlikely to have been affected.

A 95% confidence interval for the examiner level variations was ± 4.2 marks but it should be stressed that this analysis is based on raw marks and quality control procedures have not been applied. The confidence interval is ± twice the standard deviation (2.1 - from the random part), which gives ± 4.2.

Note that the centre level variation should not be interpreted as the spread of the school effects because no control has been made for the quality of the candidates at entry. The parameter estimate for centre level variation of 33.4 is quite large, but these analyses do not account for the achievement of candidates when they entered the school, and GCSE results tend to be a highly correlated with past performance. Consequently this variation is affected by prior achievement and cannot be used to make assertions about centres.

Table 9 Results of the multilevel model for History

Fixed Part

	Intercept		Candidate Sex		Examiner Sex	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
I = null	31.1	0.5	-	-	-	-
II = I +cand. sex	33.4	0.5	-3.0	0.5	-	-
III = II + exa. sex	31.1	1.5	-3.0	0.5	1.8	1.1
IV = II + masc.	33.4	0.5	-3.0	0.2	-	-
V = II +fem.	33.4	0.5	-3.0	0.2	-	-
VI = II + status	33.6	0.5	-3.0	0.2	-	-

Fixed Part (Continued)

	Masculinity		Feminity		Status	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
IV = II +masc.	-0.4	0.5				
V = II +fem.			-0.8	0.5		
VI = II + status					-1.2	1.4

Random Part

	Examiner			Centre			Candidate		
Model	Est.	s.e.	s.d.	Est.	s.e.	s.d.	Est.		s.e.
I = null	4.5	2.2	2.1	33.4	3.2	5.8	98.4	1.1	9.9
II = I +cand. sex	4.3	2.1	2.1	32.2	3.0	5.7	96.4	1.1	9.8
III = II + exa. sex	3.7	1.9	1.9	32.2	3.0	5.7	96.4	1.1	9.8
IV = II +masc.	4.3	2.1	2.1	32.2	3.0	5.7	96.4	1.1	9.8
V = II +fem.	3.9	1.9	2.0	32.2	3.1	5.7	96.4	1.1	9.8
VI = II + status	4.1	2.0	2.0	32.2	3.1	5.7	96.4	1.1	9.8