Gender and assessment: Differences, similarities and implications

James Hartley, Lucy Betts & Wayne Murray

Background
Recent changes in higher education in the UK have led to much discussion about the performance of men and women students with different methods of assessment.

Aim
To see whether or not there were differences between the marks awarded to men and women final-year psychology students as a function of the modes of assessment used.

Method
The scores obtained by 42 men and 42 women students were compared on four different methods of assessment used in their final year: a multiple-choice examination, an essay examination, a course-work essay and a project.

Results
The students obtained significantly lower scores on the multiple-choice examination than they did on the other three assessments (where they did not differ). There were no significant differences between the performance of the men and the women on these different methods of assessment when the full sample was studied. However, when the data from mature and Foundation Year students were discounted (some 20 per cent of the sample), the women performed significantly better than the men on all four measures.

Conclusions
Our students did perform differently according to the method of assessment and, to some extent, gender. Such differences suggest that it is inappropriate to pool and average the marks from different methods of assessment without first assessing whether or not there are significant differences between the marks. Furthermore, we show that different methods of standardising the results produce different distributions of students in various degree classes.

Both the study of contrasting methods of assessment, and of gender differences in the marks obtained with different methods, have a long and detailed history (see, e.g. Beard & Hartley, 1984; Brown, Bull & Pendelbury, 1997; Heywood, 2000). Nonetheless, interest seems to have been revived in both of these matters by a number of recent developments. Educationalists seem to be particularly seized by the fact that girls are now performing better than boys in almost all subject matters at school (Long, 2000), and that the proportion of women in higher education now surpasses that of men (Higher Education Statistics Agency, 2005). As a consequence we are becoming familiar with a wide variety of assertions on these topics, not all of which are soundly evidence-based. We are told, for example, that men are more likely than women to perform well on multiple-choice tests (Davies, Mangan & Telhaj, 2005), that women do better at university because of the introduction of course-work assessment (Pirie, 2001), and that men are more likely than women to be awarded first-class, third-class and pass degrees in the UK (Davies *et al.*, 2005). As Elwood (2005) points out, these are startling generalisations that do not admit to the full complexities of the situation. In all of these cases, for example, the results are affected by the ages of the students involved, the disciplines studied, the

methods of marking used, the weightings given to various examination components and how the marks are combined.

In this study we were interested to see whether or not there were differences between the marks awarded to our final-year psychology students at Keele University as a function of the modes of assessment used and of the sex of the participants. Our initial reading of this literature led us to expect that:

1. the students would have higher course-work marks than examination ones;
2. men would score more highly than women on a multiple-choice examination;
3. women would score more highly than men on both the essay-type examinations and the course-work;
4. women would have more 'good' degrees than men.

We made no predictions about the effects of different subject combinations (psychology is one of two subjects taken independently in a joint programme at Keele), or about the performance of mature students with these different methods of assessment in their final year, for the possible effects of these variables are not so clear.

We do not propose to review in detail all of the research that led us to these expectations. Instead we shall summarise some of the more recent relevant studies under three headings: (i) studies of differences between different methods of assessment; (ii) small-scale studies where gender differences have – or have not – been found; and (iii) national studies of the overall degree performance of men and women. Panel 1 summarises some of the explanations that have been given for these differences over time.

The effects of different methods of assessment

In the UK many methods of assessment are used separately and in combination in higher education. In this study we examine four of the most common ones. Multiple-choice and essay examinations are examples of direct assessments, written on the day. Course-work essays and projects/dissertations are examples of assessments of work completed over time.

Multiple-choice questions

How well students score on multiple-choice questions depends in part on the length and difficulty of the items, and on how they are

Explanations for why men do better than women

Men, it appears:

- are more able (Rudd, 1984; Irwin & Lynn, 2005)
- able to benefit from male sex bias in markers (Bradley, 1984)
- are bolder writers (Robson, Francis & Read, 2002)
- are greater risk takers (Chung & Tang, 1998)
- are more self-confident (Adams, 1986)
- are less prone to 'fear of failure' (Severiens & ten Dam, 1994)
- are less anxious about examinations (Martin, 1997)
- are more likely to adapt their approaches to learning to different contextual demands (Meyer, Dunne & Richardson, 1994)
- have more role models in that there are more male teachers in universities (Francis & Skelton, 2001)

(continued)

Panel 1: Arguments put forward in the academic literature to explain why men do better than women, or vice versa, in higher education.

(Panle 1 continued)

Explanations for why women do better than men

Women, it appears:

- have better verbal skills (Lumsden *et al*., 1987).
- are more committed to academic work (Smith, 2004).
- attend classes more assiduously (Woodfield, Jessop & MacMillan, 2006).
- collaborate with other students more when revising for examinations (Rogers, 2003; Vermunt, 2005)
- do more independent work (Woodfield, Jessop & McMillan, 2006)
- do better on course-work (see Woodfield, Earl-Novell & Solomon, 2005)
- are more likely to conform to, and less likely to be distracted from, institutional requirements (Woodfield, Jessop & McMillan, 2006)

Explanations for 'mixed effects'

Mixed results occur because:

- of disciplinary differences: males do better in the Sciences, females in the Arts and Social Sciences (Francis & Skelton, 2001)
- of differences between men and women in their ways of thinking and reasoning in different disciplines (Davies *et al*., 2005)
- of differences in the ways the sexes are distributed in different disciplines: more men get firsts and thirds because there are disproportionately more men in the sciences and the sciences award more first class degrees (Woodfield & Earl-Novell, 2006)

scored. Negative marking – where wrong answers lead to marks being taken away – and other procedures involving corrections for guessing can lead to differences in the distribution of marks on multiple-choice tests, as well as differences in their reliability. Negative marking and correction for guessing inhibits random guessing and penalises confidently held incorrect knowledge (Burton, 2001, 2002, 2004, 2005). One advantage of multiple-choice tests is that all students are set and asked to do the same questions and all are (computer) marked in the same way – thus there is no scope for biased marking here.

Essay examinations

Essay examinations are widely used in the UK, despite the fact that they are known to be unreliable as a method of assessment on at least three counts: (i) student variability on the day is not taken into account; (ii) independent examiners allocate different marks to the same scripts; and (iii) the same examiners give different marks to the same scripts if they mark them again after an interval of time. In addition there is evidence that handwriting quality, the position of the script in a series (e.g. after a run of good or after a run of poor answers) and marker-fatigue can all affect the marks given (see Hartley, 1998).

Some solutions for counteracting these problems include the use of agreed marking schemes, limiting the breadth and choice of essay topics, increasing the number of markers, marking 'blind' numbered or anonymous scripts, and computer-based marking. Coffin *et al.* (2003), Haines (2004) and Shermis, Burstein & Leacock (2006) provide useful discussions of these issues. Introducing course-work assessment is, of course, another attempt to reduce the unreliability of essay examinations by removing the problems associated with examination anxiety, etc.

Course-work assessment

Woodfield, Earl-Novell & Solomon (2005) report without qualification that, 'All (students) perform better when course-work is used as part of the assessment array'. (p. 34) – which, of course, is not literally true. It has however been shown that course-work generally increases the marks when it is included in with examination marks to arrive at an overall module mark (Bridges *et al.*, 2002; Simonite, 2003) or indeed, if it is used as the sole method of assessment (Simonite, 2003). Woodfield, Earl-Novell & Solomon also report in their study that both their men and women students preferred course-work to examinations as a method of assessment. Nonetheless, Bridges *et al*. (2002) found disciplinary differences, with higher increases in the marks when course-work was included in assessments in the Sciences and the Social Sciences than in the Arts.

Projects/Dissertations

The term 'project' covers a range of activities within which different proportions can be student-led or conceived (Cuthbert, 2001). It is commonly thought that students achieve higher marks for their projects than for their examination work largely because they are more independent and motivated in these activities. Tariq, Stefani, Butcher & Heylings (1998) provide some data to support this position but, generally speaking, there are few data on the topic. Much of the research on projects is concerned with clarifying marking schemes and procedures (e.g. Tariq *et al.*, 1998; Orsmond, Merry & Reiling, 2004).

Small scale studies of sex differences with different methods of assessment

Multiple-choice examinations

Men sometimes do better than women on multiple-choice examinations at university (e.g. see Anderson, 2002; Bridgeman & Lewis, 1994; Davies *et al.* 2005; Lumsden, Scott & Becker, 1987; Wakeford, 2003) but this result is not always found, and not all investigators check their results for sex differences (e.g. see Williams & Clark, 2004). These findings seem more common in economics, mathematics, sciences and medicine – possibly because multiple-choice questions are not commonly used as an assessment technique in the Arts. Anderson (2002) reported that women mathematics students were more likely than men to refrain from answering multiple-choice questions when they are unsure of the answers, and when there is negative marking, but Von Schrader and Ansley (2006) did not find that this with school children.

Essay examinations

Women have been found to perform better than men on essay exams in some studies (e.g. Lumsden *et al.*, 1987; Smith, 2004; Woodfield *et al.*, 2005), and there has been some discussion of the qualities of female as opposed to male essays, particularly at Oxford and Cambridge (see Woodfield *et al.*, 2005). Broadly speaking, computer-based analyses of essays written by men and women have failed to find the differences sometimes reported in hand analyses (see Hartley, 2004).

Coursework

Women, it is often thought, do better than men on course-work (Pirie, 2001), but again, this result is not always true (Elwood, 2005). Woodfield *et al.* (2005) reported that both men and women did better with course-work than with essay-type examinations. In another study, Smith (2004) reported that women did better than men on essay examinations and men did better than women on course-work in their final year in geography.

National studies of the degree performance of men and women

Most final degrees in the UK are awarded in classes – firsts, upper-seconds, lower-seconds, thirds, passes, and fails. Most honours degrees in medical subjects are not classified (though some are) and degrees awarded without honours (ordinary degrees and pass degrees) are not classified. Studies in the 1960s showed that there was a tendency for

men in the UK to receive more first-class degrees, and more thirds and passes than women. Indeed, there is still evidence for this today (see McNabb, Pal & Sloane, 2002, and Table IV in Richardson & Woodley, 2003). The reasons offered to explain this finding in the 1960s were wide-ranging (see Hartley, 1998) but today the picture is even more complicated. First of all, the proportions of men and women studying for degrees in the UK have changed dramatically, with women now occupying almost 60 per cent of the places (HESA, 2005). Second, there is a higher proportion of mature students, particularly women, in the Arts and Social Sciences. And, of course, there is much more coursework assessment. Richardson (2004) has argued that the tendency for men to be more likely to obtain first-class degrees than women is an artefact of sexist practices in the past and is now disappearing.

Woodfield and Earl-Novell (2006) provide a recent study of the distribution of degree classes obtained by men and women students in the UK. These investigators analysed the results of over 1,250,000 students who graduated between 1995 and 2002 (excluding those with unclassified degrees, combined Arts & Science degrees, and degrees in Architecture). Their results show, in effect, that any superiority shown by males in the distribution of first-class degrees could be attributed to the fact that more first-class degrees were awarded in the Sciences and that there were more men than women Science students.

The matter has not been clarified by the tendency today to pool the numbers of students obtaining 1st class and 2:1 degrees ('good' degrees) and the numbers obtaining 2:2 degrees and less ('poor' degrees). This procedure disguises the nature of any differences between men and women at the extreme ends of the distribution, especially the lower one, as it hides the proportions of students obtaining third-class degrees, passes, or fails.

However, if one looks at the results for 'good' and 'poor' degrees, the findings here do show that women do better than men in terms of 'good' degrees, and that men do worse than women in terms of 'poor' ones. Richardson and Woodley (2003), for example, reported that 57.7 per cent of the women (in their study of over 220,000 students who graduated with classified degrees in 1996) obtained 'good' degrees (compared with 51.2 per cent of the men) and that 48.8 per cent of the men obtained 'poor' degrees (compared with 42.3 per cent of the women). Richardson and Woodley further showed that these proportions were affected both by the age of the students and the discipline studied.

Panel 1 lists the main explanations frequently given for gender differences in degree performance together with some of their more recent protagonists. It is explanations such as these, together with the findings summarised above, that led us to predict what we expected to find in the present study. At this stage we were not particularly interested in assessing the evidence for or against any one of these various theoretical stances. We wanted to see what the evidence showed first before attempting to explain it. Thus we were interested to see whether or not any of the differences listed in Panel 1 held up in the specific conditions of our own Psychology course and, if they did so, then to consider the implications. Clearly, if there are large differences in the distributions of the marks obtained from different methods of assessment, and in those obtained from men and women students, then some sort of standardisation needs to be considered before the marks can be sensibly pooled to arrive at a degree class. This, of course, is not a new argument – it has been made for many years – but it may now have to be taken more seriously.

Method

Participants

In this study we used data from our departmental records to compare the performance of a sample of our men and women final year students on four different modes of assessment. Initially we examined the data from all 42 male psychology students who took their

finals in 2002 and 2003. And then, because there were far more female students than male ones in the overall sample, we matched as closely as possible these students with 42 female ones who were of a similar age (i.e. most to within one year) and who studied the same subject areas. (At Keele students study several subjects in their first year and two subjects conjointly in their second and final year: these other subjects can come from three subject areas: the Arts, the Social Sciences or the Sciences).

In their final year the median age of the participants was 21 years (range 21–51). There were, however, 10 mature students (5 in each group) (age range 28–51). In addition 5 of the traditional-entry students (and 3 of the mature students) had also completed the Foundation Year. (The Foundation Year was an extra introductory year - now no longer available - that students could choose to take *before* beginning the 3-year degree course.)

Modes of assessment

Students in the final year at Keele in 2002 and 2003 were assessed on four modules. Each module had varied components and methods of assessment within them. However four different kinds of method of assessment that were used within these modules each year were:

1. A 45-minute multiple-choice examination on neuropsychology. Here there were 80 four-choice questions and none of the fourth choices was of the 'all of the above' or 'none of the above' kind. This examination was negatively marked: correct answers were scored +3, no attempt was scored 0, and wrong answers were scored -1. This examination was one of three methods of assessment used in a final-year module entitled 'Brain and Behaviour', and it had an overall weighting of 15 per cent in determining the module grade.
2. An unseen essay-type examination. Here the students had to answer two questions from a choice of six in two hours. This examination was one of two methods of assessment used in a final-year module entitled 'Psychology and the Individual', and it had an overall weighting of 60 per cent in determining the module grade. The two essays were marked 'blind' by two members of the academic staff using a departmental essay marking guide. For this study the average mark for the two essays was recorded.
3. A coursework essay. Here all of the candidates had to write an essay (2,500 words maximum) in their own time but by a specified date as part of the requirements for completion of their particular option course (of which there were 12 each year). The essay was marked 'blind' by two members of the academic staff, one of whom was the option tutor, using a departmental essay marking guide. This essay had an overall weighting of 40 per cent in determining the module grade.
4. A project/dissertation (suggested length 5,000 words; maximum 10,000). The quantitative or qualitative research for these projects could be done individually, in pairs, or as part of a group, but they were written up individually. The projects were marked 'seen' using a departmental project marking guide by the project supervisor and by another member of the academic staff who was normally unfamiliar with these students and their work. (There were approximately 15 project supervisors each year). The project assessment had a weighting of 100 per cent for this module.

We obtained the marks for each student for each of these four methods of assessment, together with their age, sex, second subject of study, and their overall degree classification. (Note that this latter assessment included many more marks than the four studied here for psychology and that all of the psychology marks were pooled with another set of marks obtained from their second subject to arrive at this degree classification.)

We initially assessed the mean scores obtained on each of these four methods of assessment for the whole sample (N = 84).

We then repeated the analysis first with the data from the 10 mature students removed (N = 74), and then with the data from the 8 Foundation Year students removed (N = 76), and finally with the data from 15 students who belonged one or both of these overlapping groups removed (N = 69). The effects of removing the data from each subgroup was remarkably similar and somewhat startling, so we present below the data from the whole sample and then the same data again without that of the mature and the Foundation Year students.

We decided that we should carry out these separate analyses (contrary to the advice of those who suggested that we should drop these data from the mature and the Foundation Year students) because most departments have heterogeneous student populations and thus the findings have implications for us all.

Results

Table 1 shows the main results for the full and for the reduced sample. Inspection of the means for the full sample suggests that there is an effect for mode of assessment but possibly not one for sex. This was confirmed by a 2 (sex) × 4 (mode of assessment) mixed ANOVA with mode of assessment as a within-subjects factor, and with the degrees of freedom corrected for violating the assumptions of sphericity, using the Greenhouse Geisser correction method (Field, 2002). The ANOVA yielded a significant main effect for mode of assessment ($F(1.98,161.99) = 17.68$, $p < .001$, $\eta^2 = .18$). Tukey *a posteriori* tests showed that the significant difference lay between the students' performance on the multiple-choice examination and the other modes of assessment ($p < .01$), with the students achieving significantly higher scores on the other forms of assessment compared to the multiple-choice examination. There was no significant main effect of sex ($F(1,82) = 2.21$, $p > .05$, $\eta^2 = .03$) and no significant interaction between sex and mode of assessment ($F(1.98, 161.99) = 2.06$, $p > .05$, $\eta^2 = .03$).

Similar results were obtained for the reduced sample. The analysis here showed that there was a significant main effect for the mode of assessment ($F(2.12,142.24) = 18.25$, $p < .001$, $\eta^2 = 0.21$). Tukey *a posteriori* tests revealed that there were no significant differences between the performance of the students on the essay examination, the course-work essay, and the project, but that the students performed significantly better on all of these modes of assessment com-

	Full sample			Reduced sample		
	Men N = 42	Women N = 42	Total N = 84	Men N = 35	Women N = 34	Total N =69
Multiple-choice	51.0	53.4	52.2	48.6	54.8	51.7
	(15.2)	(14.5)	(14.8)	(13.7)	(14.3)	(14.2)
Exam essay	55.0	60.4	57.7	55.1	61.4	59.3
	(7.8)	(6.7)	(7.7)	(7.8)	(6.6)	(7.9)
Course-work essay	60.4	59.2	59.8	59.6	60.2	59.9
	(8.8)	(7.2)	(8.0)	(9.1)	(7.3)	(8.3)
Project	59.8	61.9	60.9	58.8	63.5	61.1
	(8.4)	(6.2)	(7.4)	(7.1)	(5.4)	(6.7)

Table 1: The means and standard deviations (in parentheses) for each method of assessment for the men and women in the full and the reduced samples.

	Methods of assessment			
	MC	EE	CE	P
Multiple-choice (MC)	–	.26[a]	.25[a]	.40[c]
Exam essay (EE)	.31[a]	–	.33[b]	.40[c]
Course-work essay (CE)	.25[a]	.31[b]	–	.39[c]
Project (P)	.39[c]	.39[c]	.34[b]	–

[a] $p < .05$, [b] $p < .01$, [c] $p < .001$

Table 2: Correlations between the methods of assessment for the full sample (top right, N = 84) and for the reduced sample (bottom left, N = 69).

pared with the multiple-choice examination ($p < .01$). In addition, with this smaller sample, there was also a significant main effect of sex ($F(1,67) = 8.84$, $p < .01$, $\eta^2 = .12$). Here the women scored significantly higher on all of the assessments compared to the men. There was, however, no interaction between mode of assessment and sex ($F(2.12,142.24) = 1.82$, $p > .05$, $\eta^2 = .03$).

Table 2 shows the inter-correlations between the marks obtained on the different methods of assessment for the full and for the reduced sample. It can be seen that these correlations, whilst statistically significant, are not particularly high. This suggests that these modes of assessment are measuring different skills (or are unreliable). It also supports the notion that marks on different measures need to be adjusted or standardised in some way if they are going to be pooled to give a single overall score. Separate analyses of these inter-correlations for the men and the women students in both the full and the reduced samples were computed but the results did not differ significantly from those shown in Table 2 and are thus not reported here.

Examination marks versus course-work marks

The data provided in Table 1 show that the students obtained significantly lower scores on the multiple-choice examination but did not perform significantly differently on the three other measures. Accordingly it did not seem reasonable to pool together and compare the results from the two examination measures and the two course-work ones separately as we had originally intended.

'Good' versus 'poor' degrees

Table 3 shows the results that we obtained when the number of students obtaining 'good' degrees (1sts and 2:1s combined) was compared with the number obtaining 'poor' degrees (2:2s and below) for both the full and the reduced samples. These data suggest that there are no significant

		Full sample		Reduced sample	
		Men	Women	Men	Women
'Good' degrees	N	22	28	18	27
	%	(52)	(67)	(51)	(79)
'Poor'degrees	N	20	14	17	7
	%	(48)	(33)	(49)	(21)

Table 3: The number and percentage of 'good' versus 'poor' degrees according to gender for the full and the reduced sample.

differences between the percentages of the men and the women in each of these categories for the full sample, and this was confirmed by statistical analysis ($\chi^2(1) = 1.78$, $p > .05$). However, as before, significant results were obtained with the reduced sample. Specifically, the women were awarded more 'good' degrees and fewer 'poor' degrees compared to the men ($\chi^2(1) = 5.95$, $p < .05$).

Effects of subject combinations

Table 4 shows the results for the four methods of assessment for the full and the reduced samples when the students were grouped according to whether or not their other subject fell into the main disciplinary areas of the Arts, the Social Sciences or the Sciences. The means for both samples were analysed using a 3 (subject combination) × 4 (mode of assessment) mixed ANOVA with mode of assessment as a within-subjects factor and the degrees of freedom corrected for violating the assumption of sphericity (Field, 2002). For the full sample there were no significant main effects but there was a significant interaction between mode of assessment and subject combination ($F(4.16,168.50) = 2.70$, $p < .05$, $\eta^2 = .06$). Simple comparisons revealed that there were significant differences between the performance of the Arts and Social Science students across the mode of assessment ($F(2, 168.50) = 10.58$, $p < .01$ and $F(2.17,168.50) = 10.87$, $p < .01$ respectively) but that there were no significant differences in how the Natural Science students performed in this respect ($F(1.80, 168.50) = 2.39$, $p > .05$). Tukey *a posteriori* tests also revealed that there were significant differences between the students' performance on the multiple-choice examination and all of the other modes of assessment with the Natural Science students performing significantly better on the multiple-choice examination ($p < .01$).

A similar mixed ANOVA was used to compare the means for the reduced sample. When the data from the mature and the Foundation Year students were excluded there was a significant main effect for mode of assessment ($F(2.20, 145.29) = 13.75$, $\eta^2 = .17$). Tukey *a posteriori* tests revealed that students performed worse on the multiple-choice examination than they did on the other modes of assessment ($p < .01$). However, for this reduced sample, there was no significant main effect of subject combination ($F(2,66) = .34$, $p > .05$, $\eta^2 = .01$) and no significant interaction between subject combination and mode of assessment ($F(4.40. 145.29) = 2.18$, $p > .05$ $\eta^2 = .06$).

	Other degree area combined with Psychology					
	Full sample			Reduced sample		
	Arts N = 18	Social N = 50	Natural N = 16	Arts N = 17	Social N = 38	Natural N =14
Multiple-choice	47.3	52.0	58.2	48.2	50.9	58.1
	(13.3)	(13.9)	(18.1)	(13.1)	(11.9)	(19.4)
Exam essay	58.8	58.3	54.4	59.2	58.9	55.4
	(7.2)	(6.5)	(10.8)	(7.3)	(6.6)	(11.1)
Course-work essay	59.4	59.7	60.7	59.9	59.6	60.9
	(8.0)	(9.0)	(5.9)	(6.9)	(9.5)	(6.2)
Project	60.8	60.9	60.7	61.4	60.8	61.7
	(5.0)	(7.9)	(8.4)	(4.7)	(7.0)	(8.4)

Table 4: The means and standard deviations (in parentheses) for each method of assessment for students with different subject area combinations for the full and the reduced samples.

Discussion

The main results of this study show *for the full sample* that:

1. There were statistically significant differences between the examination marks obtained on the different modes of assessment. Notably, the students scored significantly lower on the multiple-choice test than they did on the other modes of assessment.
2. There was no significant difference between the performance of the men and the women on these different modes of assessment.
3. The scores obtained on the different methods of assessment were not highly correlated.
4. Subject combinations played a small part in the results, in that students doing Psychology with another Natural Science subject performed better than students doing Psychology with an Arts or Social Science subject on the multiple-choice test.

However, *when the data from the mature and the Foundation Year students were removed*, significant differences were found between the performances of the men and the women on all of the measures employed, with the women outperforming the men in each case.

These results complement the findings discussed in the Introduction and reported in Panel 1. Thus:

- The students did perform differently on the multiple-choice examination compared with the other methods of assessment. The lower marks obtained here could reflect the nature of the examination subject-matter (neuropsychology) but they could also possibly be attributed to incorrect and random guessing which was penalised (see Burton, 2002). Also of interest here was the fact that the standard deviations of the scores in the multiple-choice examination in this study were almost double those obtained with the other methods of assessment (unlike those found by Simonite (2003) where they were much the same). An additional problematic issue here (on which we had no data) was whether or not there were differences in the number of students who did not attempt particular questions. Finally, we should note here that, with the reduced sample, the women students scored significantly better than the men on the multiple-choice examination, contrary to the findings reported in the Introduction.
- The students did *not* obtain higher marks on *both* of the course-work components compared with *both* of the examination components of their assessments, as they performed better on three of these assessments and worse on one. These findings are therefore not in line with the findings reported in the Introduction - although they partly support those of Smith (2004).
- In the reduced sample the women students did better than the men on *all* of the modes of assessment compared here. Furthermore, in the reduced sample, the women students obtained better overall degrees than did the men, as reported by Richardson and Woodley (2003). However, as our students were joint-honours ones, it was not possible to relate these findings to those reported by Woodfield & Earl-Novell (2006).
- With the full sample, some disciplinary differences were shown when the marks were combined with those from other disciplines (but *not* along the lines reported by Bridges *et al.*, 2002). Our students performed differently in accordance with their subject combina-tions on the multiple-choice test and they did not perform differently on the course-work components of the assess-ments when they were grouped in different subject combinations as reported by Bridges *et al.* (2002).

The need to standardise marks

When the marks on different tests do not correlate well, and when there are differences between means and standard devia-

tions obtained with different methods of assessment, then the marks need to be standardised in some way in order to arrive at a fair, final overall assessment. But what method is appropriate?

Brown, Bull and Pendelbury (1997) present two dramatic tables to show what happens to the students' average marks and their ranking in the list (i) when two markers use different ranges in marking their assessments (p.239), and (ii) when the original raw marks are standardised (p.240). In both of these (selected) illustrations, some students who come out top on one measure come out bottom on another.

Heywood (2000) suggests that it is inappropriate to standardise marks when the assessments measure different things (leading to the notion of separate marks for different skills and portfolio assessment). However, Heywood also argues that it is not worthwhile to bother with standardisation when the marks of the examiners are consistently close. If one or two measures seem out of alignment, though, he suggests that it is best to standardise around the one measure that seems to be the most central.

Both Heywood and Brown *et al.* are sanguine about the effects of pooling sets of marks from different disciplines: they simply point to the difficulties that arise. Certainly the results from the study of Bridges *et al.* (2002) are illuminating here. Bridges *et al.* reported that there were wide differences between the marks given to course-work assessments in different disciplines, and they showed how some students who combined different disciplines in their degrees could be advantaged or disadvantaged by this. A study by Yorke *et al*. (2004) similarly showed that different algorithms used by different universities to arrive at final degree classifications produced different results.

Heywood and Brown *et al.* are equally sanguine about the role played by external examiners in this process. Today many external examiners only have access to the marks obtained by students on individual modules and never get to see a student's entire profile. External examiners can change the marks of individuals (and of whole courses) in different disciplines without reference to the effects of this on other individuals (or courses).

The effects of standardisation

In order to illustrate the effects of using different methods of standardisation we present here the results from applying two such methods to the data shown in Table 1 for the full sample. With Method 1 we transformed the data from the four different methods of assessment so that each one had a mean of 60 and a standard deviation of 10, as this distribution typically reflects the average of our normal non-standardised results at Keele. With Method 2, we standardised the data on to the mean and standard deviation of the original essay exam scores (m = 58, s.d. 7.70) as these data represent the most central measure (as recommended by Heywood). Finally, we calculated the number of students that would fall into the various degree categories (first 70–100; upper-second 60–69; lower-second 50–59; third 40–49; and pass 35–39) using these two different methods. Table 5 shows the results obtained: a) without standardisation, (b) standardised using method 1, and (c) standardised using method 2. It is clear from this table that different methods produce different results.

Table 5 is provided for illustrative purposes only and not to recommend a particular solution. Indeed, the reader needs to be reminded

	Degree Class				
	1st	2i	2.2	3rd	Pass
Non-standardised	3	24	50	;6	1
Standardised (Method 1)	9	35	35	4	1
Standardised (Method 2)	1	26	52	5	0

Table 5: The numbers of students in different degree classifications depending upon the method of standardisation.

at this point that the data we have discussed above is only part of the data that were available to the department in arriving at the students' overall marks. Nonetheless, it is clear with these current data that these two different methods of standardisation produce different distributions of students falling into the various degree classes. These differences may be small but, for students on a borderline, they are important.

Summary

With our heterogeneous full sample of students (N = 84) we did not find any differences between the performance of men and women on four components taken from a final year examination diet. We did find, however, that there were some differences on the mean scores of these components – with our students doing significantly less well on a multiple-choice examination. Further, students combining Psychology with another Natural Science subject did significantly better on the multiple-choice test (on which was on neuropsychology) than did students combining Psychology with another Arts or Social Science subject. However, with our more homogeneous reduced sample (N = 69), although the main findings concerning the modes of examination was repeated, women performed significantly better than men on all four of examination components.

These results led us to suggest that, in arriving at an overall *examination* score for each student, we needed to standardise the results from different examinations. We compared two ways of doing this, and found that these different ways produced different numbers of students in each degree class. One implication of this is that single-honours departments need to decide on what they believe to be an appropriate way of standardising results for their students. Another is that this matter is far more complicated for joint-degree courses.

Acknowledgements

We are grateful to John Richardson and Richard Burton, and two anonymous referees, for helpful comments on an earlier version of this paper.

References

Adams, R. (1986). Some contributions to sex differences in scholastic aptitude score. *Studies in Educational Evaluation, 12,* 267–274. (Cited in Davies, *et al*, 2005).

Anderson, J. (2002). Gender-related differences on open and closed assessment tasks. *International Journal of Mathematical Education in Science and Technology, 33*(4), 495–503.

Beard, R., & Hartley, J. (1984). *Teaching and learning in higher education* (fourth edition). London: Paul Chapman.

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

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*(1), 37–50.

Bridges, P., Cooper, A., Evanson, P., Haines, C., Jenkins, D., Scurry, D., Woolf, H., & Yorke, M. (2002). Coursework marks high: Examination marks low: Discuss. *Assessment and Evaluation in Higher Education, 27*(1), 35–48.

Brown, G., Bull, J., & Pendlebury M. (1997). *Assessing Student learning in higher education.* London: Routledge.

Burton, R.F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: Question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26*(1), 41–50.

Burton, R.F. (2002). Misinformation, partial knowledge and guessing in true/false tests. *Medical Information, 36*(9), 805–811.

Burton, R.F. (2004). Multiple-choice and true/false tests: Reliability measures and some implications of negative marking. *Assessment and Evaluation in Higher Education, 29*(5), 585–595.

Burton, R.F. (2005). Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment and Evaluation in Higher Education, 30*(1), 65–72.

Chung, J., & Tang, K. (1998). Inherent gender differences as an explanation of the effect of instructor gender on accounting students' performance. In B. Black & N. Stanley (Eds.), *Teaching and learning forum: Teaching and learning in changing times.*

Perth: University of Australia. (Cited in Davies *et al*., 2005).

Coffin, C., Curry, M. J., Goodman, S., Hewings, A., Lillis, T., & Savage, J. (2003). *Teaching academic writing: A toolkit for higher education*. London: Routledge.

Cuthbert, K. (2001). Independent study and project work: Continuities and discontinuities. *Teaching in Higher Education, 6*(1), 69–84.

Davies, P., Mangan, J., & Telhaj, S. (2005). Bold, reckless and adaptable? Explaining gender differences in economic thinking and attitudes. *British Educational Research Journal, 31*(1), 29–48.

Elwood, J. (2005). Gender and achievement: What have exams got to do with it? *Oxford Review of Education, 31*(3), 373–393.

Field, A. (2002). *Discovering statistics using SPSS for Windows*. London: Sage.

Francis, B., & Skelton, C. (Eds.). (2001). *Investigating gender: Contemporary perspectives in education*. Buckingham: Open University Press.

Haines, C. (2004). *Assessing students' written work: Marking essays and reports*. London: RoutledgeFalmer.

Hartley, J. (1998). *Learning and studying: A research perspective*. London: Routledge.

Hartley, J. (2004). Is academic writing masculine? *Higher Education Review, 37*(2), 53–62.

Heywood, J. (2000). *Assessment in higher education* (second edition). London: Jessica Kingsley.

Higher Education Statistics Agency (2005). Higher education statistics: http://www.hero.ac.uk/uk/inside_he/higher_education_statisitics 6763.cfm (Accessed 29/06/2005).

Irwin, P., & Lynn, R. (2005). Sex differences in means and variability on the Progressive Matrices in university students. *British Journal of Psychology, 96*(4), 505–524.

Long, M. (2000). *The psychology of education*. London: RoutledgeFalmer.

Lumsden, K.G., Scott, A., & Becker, W.E. (1987). The economics student re-examined: Male-female differences in comprehension. *Journal of Economic Education, 18*(4), 365–374.

Martin, M. (1997). Emotional and cognitive effects of examination proximity in female and male students. *Oxford Review of Education, 23*(4), 479–486.

McNabb, R., Pal, S. & Sloane, P. (2002). Gender differences in educational attainment: The case of university students in England and Wales. *Economica, 69*, 481–503.

Meyer J., Dunne, T., & Richardson, J. (1994). A gender comparison of contexualised study behaviour in higher education. *Higher Education, 27*, 469–487.

Orsmond, P., Merry, S. & Reiling, K. (2004). Undergraduate project work: Can directed tutor support enhance skills development? *Assessment & Evaluation in Higher Education, 29*(5), 625–641.

Pirie, M. (2001). How exams are fixed in favour of girls. *The Spectator,* (January 20). (Cited by Woodfield, Earl-Novell & Solomon.)

Richardson, J.T.E. (2004). Contingent degree performance. *The Psychologist, 17*(6), 323–324.

Richardson, J.T.E., & Woodley, A. (2003). Another look at the role of age, gender and subject as predictors of academic attainment in higher education. *Studies in Higher Education, 28*(4), 475–493.

Robson, J., Francis, B., & Read, B. (2002). Writes of passage: Stylistic features of male and female undergraduate history essays. *Journal of Further and Higher Education, 26*(4), 351–362.

Rogers, L. (2003). Gender differences in approaches to studying for GCSE among Year 10 pupils. *Psychology of Education Review, 27*(1), 18–27.

Rudd, E. (1984). A comparison between the result achieved by men and women studying for first degrees in British Universities. *Studies in Higher Education, 9*(1), 47–57.

Severiens, S., & ten Dam, G. (1994). A multi-level meta-analysis of gender differences in learning orientations. *British Journal of Educational Psychology, 68*, 595–608.

Shermis, M. D., Burstein, J., & Leacock, C. (2006). Applications of computers in assessment and analysis of writing. In C.A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research,* pp. 403–416. New York: Guilford Press.

Simonite, V. (2003). The impact of coursework assessment on degree classifications and the performance of individual students. *Assessment and Evaluation in Higher Education, 28*(5), 454–470.

Smith, F. (2004). 'It's not all about grades': Accounting for gendered degree results in Geography at Brunel University. *Journal of Geography in Higher Education, 28*(2), 167–178.

Tariq, V. N., Stefani, L. A. J., Butcher, A. C., & Heylings, D. J. A. (1998). Developing a new approach to the assessment of project work. *Assessment & Evaluation in Higher Education, 23*(3), 221–240.

Vermunt, J.D. (2005). Relations between student learning patterns and personal and contextual factors and academic performance. *Higher Education, 49*, 205–234.

Von Schrader, S., & Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education, 19*(1), 41–65.

Wakeford, R. (2003). Principles of student assessment. In H. Fry, S. Ketteridge & S. Marshall (Eds.). *A handbook for teaching and learning in higher education* (second edition). London: Kogan Page (pp.42–61).

Williams, R.L., & Clark L. (2004). College students' ratings of student effort, student ability and teacher input as correlates of student performance on multiple-choice exams. *Educational Research, 46*(3), 229–239.

Woodfield, R., & Earl-Novell, S. (2006). An assessment of the extent to which subject variation between the Arts and the Sciences in relation to the award of first-class degrees can explain the 'gender gap' in UK universities. *British Journal of Sociology of Education, 27*(3), 355–372.

Woodfield, R., Earl-Novell, S., & Solomon, L. (2005). Gender and mode of assessment at university: Should we assume female students are better suited to coursework and males to unseen examinations? *Assessment and Evaluation in Higher Education, 30*(1), 33–48.

Woodfield, R., Jessop, D., & McMillan, L. (2006). Gender differences in undergraduate attendance rates. *Studies in Higher Education, 31*(1), 1–22.

Yorke, M., Barnett, G., Evanson, P., Haines, C., Jenkins, D., Knight, P., Scurry, D., Stowell, M., & Woolf, H. (2004). Some effects of the award algorithm on honours degree classifications in UK higher education. *Assessment & Evaluation in Higher Education, 29*(4), 401–413.