Investigating male bias in multiple choice questions: contrasting formative and summative settings

OPEN ACCESS
IOP Publishing
European Journal of Physics
Eur. J. Phys. **39** (2018) 055704 (7pp)
https://doi.org/10.1088/1361-6404/aad169

Investigating male bias in multiple choice questions: contrasting formative and summative settings

H Hedgeland, H Dawkins and S Jordan

School of Physical Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, United Kingdom

E-mail: holly.hedgeland@open.ac.uk

Received 22 April 2018, revised 25 June 2018
Accepted for publication 5 July 2018
Published 23 July 2018

Abstract

Previous studies have claimed that male advantage may arise from multiple choice question (MCQ) types; we have made a detailed evaluation of this hypothesis, finding limited evidence that female students are disadvantaged by MCQs in summative assessment. Additionally, we find no significant evidence of a gender gap around the use of multiple choice-type questions, including variants such as multiple response questions, in formative assessment. Our findings suggest that the use of a MCQ format is not a significant factor in the gender gap in assessment.

Keywords: multiple choice, gender differences, assessment, women in physics

(Some figures may appear in colour only in the online journal)

1. Introduction and context

A gender gap in attainment in physics is frequently observed and has been noted in situations employing multiple choice-type questions, such as concept inventories (Docktor and Heller 2008, Kost-Smith *et al* 2010, Bates *et al* 2013, Madsen *et al* 2013, Traxler *et al* 2018), compounded by the widely-held belief of general male advantage in multiple choice (Ben-Shakhar and Sinai 1991, Gipps and Murphy 1994, Arthur and Everaert 2012, Wilson *et al* 2016). At The Open University, UK, we have seen a persistent trend of higher attainment

by males in second level physics and astronomy modules (at Level 5 in the Framework for Higher Education Qualifications). Previously, we considered the effect of written exam question scaffolding on this gap (Dawkins *et al* 2017). However, the use of multiple choice-type questions, alongside others, in both formative and summative assessment in these modules also leaves us well-placed to address the research question of whether there is evidence to support the assertion that the gender gap is exacerbated by the use of these question types.

The Open University offers open access distance learning courses with a substantial online component. A total of 360 credits are required for a degree, and students are encouraged to progress through a defined qualification pathway in order to provide themselves with adequate mathematical preparation prior to attempting the second and third level physical sciences. There are no formal pre-requisites at entrance and our students have a diverse range of educational backgrounds and motivations for their studies. We have a significant number of part-time students who are studying at a later stage of life than in most conventional universities and who contribute to a demographically diverse student population. However, despite these differences, we see a similar gender gap in attainment to that noted more widely within the sector.

In this paper we examine the role of the use of multiple choice questions (MCQs) in this attainment gap, firstly by considering their use in a summative setting. Our 60 credit core physics module at Level 2 (S207) and the two 30 credit astronomy modules (S282 and S283) are the first opportunity for the study of physics or astronomy as individual disciplines and present key topics at an introductory to intermediate level. The physics course covers a substantial fraction of the Core of Physics as defined by the UK Institute of Physics, including material on all the major topics. The astronomy modules form an introduction to the Sun, stars, galaxies and cosmology (S282) and planetary science and astrobiology (S283). Data collected over three presentations of each module (totalling 1270 students on S207 (24% female), 712 students on S282 (30% female) and 601 students on S283 (32% female)), allow us to compare the performances of male and female students in the multiple choice computer-marked section of the exam with those in the constructed response section.

Additionally, we make use of interactive computer-marked assignments (iCMAs), short problems requiring numerical open responses or selected responses (such as multiple choice), that are used in formative assessment in Level 2 physics. In this study we analyse iCMA responses from a total of 1411 students (75% male; 25% female) to identify any gender bias and its variation by question type, to allow us to explore the difference between the use of multiple choice-type questions in formative and summative contexts.

2. Quantitative analysis and interpretation

2.1. Summative assessment

In table 1 we present the mean percentage scores in the MCQ and written sections of the end of course examination of the male and female students from each of three cohorts of the Level 2 physics and astronomy modules. In the majority of cohorts, we see both males and females achieving higher marks in the MCQ section of the exam. We also see a tendency for the increase in the scores of the males in the MCQ section to be greater than that of the females, and the sixth column of table 1 shows the female score difference subtracted from that of the males. Positive values indicate situations where the male score has increased by more, or decreased by less, than the female one in the multiple choice section. Although this is

Table 1. Differing attainment in MCQ and written sections of summative assessment; the gender difference in the variation between scores in the MCQ and written sections (MCQ—*written* scores for the females subtracted from MCQ—*written* scores for the males); probability of male and female scores coming from distributions with the same mean. Scores are given as percentages. The distributions of the mean scores are typically truncated normal, with a standard deviation of around 20.

Cohort	Mean MCQ score		Mean written score		Gender difference	Probability same	
	Males	Females	Males	Females		MCQ	Written
S207 2011–2	74.6	69.1	59.1	56.3	2.7	0.020	0.202
S207 2012–3	75.6	73.9	47.9	45.0	−1.2	0.454	0.234
S207 2013–4	74.2	69.6	58.8	54.9	0.7	0.074	0.187
S282 2013	64.1	57.5	55.8	51.7	2.5	0.014	0.115
S282 2013–4	64.5	55.0	54.4	49.2	4.3	0.002	0.089
S282 2014–5	62.2	58.7	65.6	57.7	−4.4	0.237	0.037
S283 2013–4	69.9	65.1	55.9	53.7	2.6	0.072	0.398
S283 2014–5	57.1	51.0	59.7	59.5	5.9	0.013	0.933
S283 2015–6	66.7	63.4	59.0	58.7	3.0	0.115	0.998

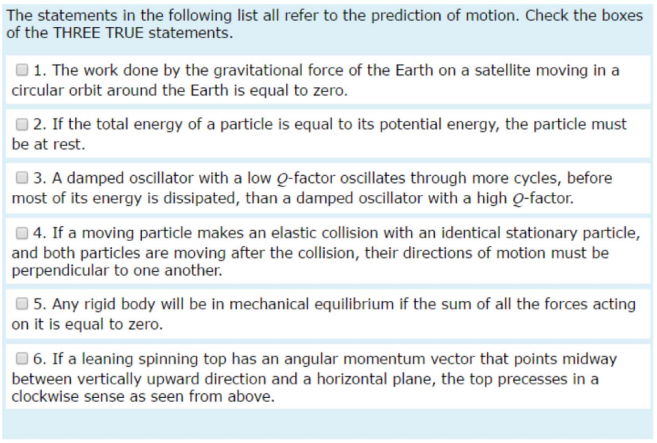
indicative of possible male bias, it is not conclusive as in both situations bias will be convolved with any difference in ability between the male and female cohorts.

To give us an indication of the magnitude of this effect, we consider the male and female results from the two summative assessment sections separately. By carrying out a Welch's t-test, we find the probability that the male and female scores on each section represent no real difference between the mean scores in that section. The probabilities are presented in the final two columns of table 1. A low probability of the true means of the male and female scores in the section being the same could be caused either by gender bias within the section, or by a difference in ability between the male and female students. Marked differences between the two sections are hence of interest as these indicate factors other than ability are at play.

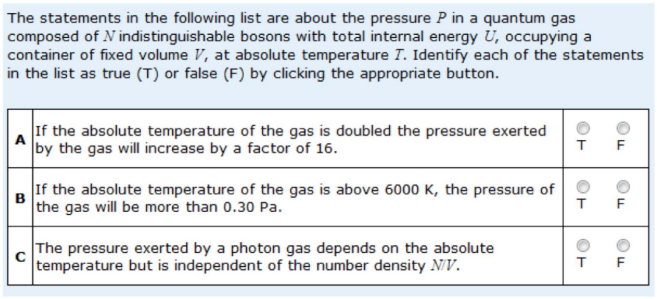
We test to a significance level of 0.05, with a Bonferroni correction applied for nine tests so that each cohort is tested at 0.0056. Only S282 in 2013–4 shows a significant difference between the means, in the MCQ section of the assessment. However, the 0.089 probability of the same mean in the written section implies ability cannot be discounted here. The opposite effect is seen in the 2014–5, with $p = 0.013$ in the MCQ and 0.933 in the written section, which is suggestive of bias in the MCQ section for that particular paper, although not to a level that is statistically significant. Comparison of these figures for all cohorts shows that there is limited evidence of male bias in MCQ sections of specific examination papers but provides no support for consistent male advantage. Overall, the variation between modules and cohorts is notable and no consistent significant effect is seen around the use of MCQs in the summative setting.

2.2. Formative assessment

We consider now iCMA responses from three Level 2 physics cohorts, overlapping with the summative cohorts (through 2012–3 and 2013–4). (iCMAs are used only in the physics modules and not in astronomy.) In the iCMAs, we identified 15 of the 56 questions as taking multiple choice formats. Of these, eight were multiple response questions (MRQs), four were text-based MCQs, or questions containing such an MCQ element, one was a graph-based



(a) Typical multiple response format



(b) Example of true/false format

Figure 1. Illustrating two of the multiple choice question types used in the formative assessments.

MCQ and two were in a true/false choice format. Two of these question types are shown in figure 1 to illustrate the variation beyond the basic MCQs that are used in the summative setting. Given these subtleties in question presentation, we wished to evaluate whether individual questions demonstrated any significant male bias, while accounting for student ability. Students were divided into strata by ability, defined by their overall performance on the full iCMA question set. We then calculated a Mantel–Haenszel alpha for each question, which finds the ratio of the success probabilities between the groups of the male and female students by evaluating

$$\alpha^{*}_{\mathrm{MH}} = -2.35 \ln\left(\frac{\sum_i m_i^1 f_i^0 \ / N_i}{\sum_i m_i^0 f_i^1 \ / N_i}\right), \tag{1}$$

where m_i^1 is the number of males in the the ith stratum answering the question correctly and m_i^0, the number incorrectly, and similarly f_i^1 and f_i^0 for the females, such that the total number of students in the stratum is $N_i = m_i^1 + m_i^0 + f_i^1 + f_i^0$ (Osterlind and Everson 2009). (Twenty strata were used in our analysis.) We used the logarithmic transform variant of the measure (which includes the numerical prefactor) as this produces a symmetrical scale around zero. A negative value of α^{*}_{MH} indicates a male bias, with a greater probability of a male student answering the question correctly than a female student of the same ability.

Table 2. Question details, α^*_{MH} and p values for the 15 multiple choice-type questions used in the iCMAs.

Type	Topic	α^*_{MH}			p		
		2012–3	2013–4	2014–5	2012–3	2013–4	2014–5
MRQ	Mechanics	−1.30	−1.27	−2.23	0.71	0.91	0.14
MRQ	Mechanics	−1.32	0.39	−0.15	0.97	0.04	0.69
MRQ	Electromagnetism	−2.08	−0.33	−2.11	0.35	0.60	0.60
MRQ	Electromagnetism	−0.82	−0.06	−0.32	0.87	0.15	0.17
MRQ	Electromagnetism	−1.30	−1.92	−1.88	0.76	0.96	0.92
Graph MCQ	Waves	−0.54	−1.94	−2.04	0.20	0.76	0.27
MCQ	Optics	−3.38	−2.31	−3.00	0.49	0.56	0.82
MCQ	Quantum mechanics	−3.08	−1.67	−2.75	0.50	0.09	0.61
MRQ	Quantum mechanics	0.14	−2.57	−0.58	0.02	0.32	0.13
MRQ	Thermodynamics	−0.75	−1.10	−0.15	0.88	0.93	0.55
MCQ	Thermodynamics	0.09	0.04	1.51	0.53	0.91	0.06
True/False	Thermodynamics	−0.82	0.33	−1.68	0.42	0.56	0.58
MRQ	Condensed matter	−1.15	−1.38	−0.70	0.28	0.78	0.15
MRQ	Condensed matter	−0.62	−0.88	−0.48	0.69	0.34	0.66
True/False	Condensed matter	−1.71	−1.05	0.48	0.36	0.87	0.18

Conversely, a positive value suggests a female bias, with values of $|\alpha^*_{MH}| \geqslant 1$ deemed to be potentially significant in either respect. Using a chi-squared distribution, each alpha value is also tested against the null hypothesis that the odds ratio is equal to one at each stratum, with an alternative hypothesis that at least one odds ratio is different from unity. A question is deemed to have significant bias if $p \leqslant .05$, in addition to $|\alpha^*_{MH}| \geqslant 1$ (Zwick 2012). Applying the Bonferroni correction here would suggest $p \leqslant 0.001\,1$ as appropriate to determine significance.

Of the 15 questions in multiple choice formats, none are observed to have a significant male or female bias. Interestingly, the two questions with the lowest values of p (0.02 and 0.04) demonstrate a slight female bias (with corresponding α^*_{MH} values of 0.14 and 0.39). Both were MRQs, one on the topic of mechanics (illustrated in figure 1(a)) and the other, quantum physics. The next lowest ($p = 0.06$), showed a stronger female bias ($\alpha^*_{MH} = 1.5$) in one cohort; this question contained a substantial MCQ element and covered aspects of thermodynamics. The full set of α^*_{MH} and p values is shown in table 2. It is particularly interesting to note that the S207 2012–3 and 2013–4 cohorts show no significant evidence of bias, in parallel with the behaviour they demonstrated in the summative setting. In total, we find that there is no evidence to suggest a female disadvantage owing to multiple choice-type questions within a formative setting.

3. Future work

Our finding of no significant male bias around the use of multiple choice-type questions in either a formative or summative setting is not in agreement with a number of other studies concerning the use of MCQs in physics assessment. Hazel *et al* (1997) suggest a MCQ attainment gap in physics and call for the use of the question type only in a diagnostic setting where common misconceptions can be employed as distractors. More recently, Wilson *et al* (2016) continue to note a gender gap around the use of MCQs in the competitive setting of

Olympiad team selection. Recent work by Kost-Smith *et al* (2010) suggests that the difference seen on conceptual surveys, usually presented in MCQ format, could be ascribed to the accumulation of small gender differences over time. If we do not seek to question the validity of this body of work, we must then look to what might be different about the questions, or approaches, we have used.

In the formative data, we found suggestion of occasional female advantage. As we described, and illustrated in figure 1, the iCMA questions are not conventional MCQs, but can involve reading a reasonable quantity of text, which has been noted to favour female students and, for example, is one of the factors used by Wilson *et al* (2016) in analysis of their MCQ question set. The structure of our formative assessment also permits multiple attempts, with limited feedback. Whilst not all students wish to engage with the questions in this way, its availability may provide the student with a more positive experience of these question types than if they had only ever encountered them in a competitive, summative setting, and could hence be related to the idea of gender difference reflecting cumulative effects. Exploring this potential connection to the student experience and the students' wider background is of interest for future work.

4. Conclusions

We find equivocal evidence of an increased gain for males over females when MCQs are used in summative assessment. Our findings highlight the need for careful use of this question type in examinations but do not support the view that it is intrinsically problematic. When used in formative assessment, we found no evidence of male bias across a variety of MCQ formats and topics in physics covering mechanics, optics and electromagnetism, thermodynamics, quantum mechanics and solid-state physics.

Acknowledgments

We gratefully acknowledge assistance from Richard Jordan and financial support from the Open University's eSTEeM centre for STEM pedagogy.

ORCID iDs

H Hedgeland https://orcid.org/0000-0003-3703-7942
S Jordan https://orcid.org/0000-0003-0770-1443

References

Arthur N and Everaert P 2012 Gender and performance in accounting examinations: exploring the impact of examination format *Account. Educ.* **21** 471–87

Bates S, Donnelly R, MacPhee C, Sands D, Birch M and Walet N R 2013 Gender differences in conceptual understanding of newtonian mechanics: a UK cross-institution comparison *Eur. J. Phys.* **34** 421

Ben-Shakhar G and Sinai Y 1991 Gender differences in multiple-choice tests: the role of differential guessing tendencies *J. Educ. Meas.* **28** 23–35

Dawkins H, Hedgeland H and Jordan S 2017 Impact of scaffolding and question structure on the gender gap *Phys. Rev. Phys. Educ. Res.* **13** 020117

Docktor J and Heller K 2008 Gender differences in both force concept inventory and introductory physics performance *AIP Conf. Proc.* vol 1064 pp 159–62

Gipps C V and Murphy P 1994 *A Fair Test? Assessment, Achievement, and Equity* (Buckingham, Philadelphia: Open University Press)

Hazel E, Logan P and Gallagher P 1997 Equitable assessment of students in physics: importance of gender and language background *Int. J. Sci. Educ.* **19** 381–92

Kost-Smith L E, Pollock S J and Finkelstein N D 2010 Gender disparities in second-semester college physics: The incremental effects of a 'smog of bias' *Phys. Rev. Spec. Top.—Phys. Educ. Res.* **6** 020112

Madsen A, McKagan S B and Sayre E C 2013 Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Phys. Rev. Spec. Top.—Phys. Educ. Res.* **9** 020121

Osterlind S J and Everson H T 2009 *Differential Item Functioning, Vol 161 of Quantitative Applications in the Social Sciences* (Thousand Oaks, CA: Sage Publications)

Traxler A, Henderson R, Stewart J, Stewart G, Papak A and Lindell R 2018 Gender fairness within the force concept inventory *Phys. Rev. Phys. Educ. Res.* **14** 010103

Wilson K, Low D, Verdon M and Verdon A 2016 Differences in gender performance on competitive physics selection tests *Phys. Rev. Phys. Educ. Res.* **12** 020111

Zwick R 2012 A review of ets differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement *ETS Res. Rep. Ser.* i–30