Investigating Male Bias in Multiple Choice Questions: Contrasting Formative and Summative Settings

How to cite:

For guidance on citations see FAQs.

© 2018 European Physical Society

Version: Accepted Manuscript

Link(s) to article on publisher’s website:

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.
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, UK
E-mail: holly.hedgeland@open.ac.uk

Abstract. Previous studies have claimed that male advantage may arise from multiple choice question types; we have made a detailed evaluation of this hypothesis, finding limited evidence that female students are disadvantaged by multiple choice questions 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 multiple choice question format is not a significant factor in the gender gap in assessment.

PACS numbers: 01.40.-d
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 (Bates, Donnelly, MacPhee, Sands, Birch & Walet 2013, Kost-Smith, Pollock & Finkelstein 2010, Docktor & Heller 2008, Madsen, McKagan & Sayre 2013, Traxler, Henderson, Stewart, Stewart, Papak & Lindell 2018), compounded by the widely-held belief of general male advantage in multiple choice (Ben-Shakhar & Sinai 1991, Arthur & Everaert 2012, Wilson, Low, Verdon & Verdon 2016, Gipps & Murphy 1994). 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, Hedgeland & Jordan 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
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.

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

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 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, 8 were multiple response questions (MRQs), 4 were text-based MCQs, or questions containing such an MCQ element, 1 was a graph-based MCQ and 2 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:

\[
\alpha_{MH}^* = -2.35 \ln \left( \frac{\sum_i m_i^1 f_i^0 / N_i}{\sum_i m_i^0 f_i^1 / N_i} \right). 
\]  

where \( m_i^1 \) is the number of males in the the \( i \)th 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 & 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 \( \alpha_{MH}^* \) indicates a male bias, with a greater probability of a male student answering the question correctly than a female student of the same ability. Conversely, a positive value suggests a female bias,
with values of $|\alpha_{MH}^*| \geq 1$ 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 $p \leq .05$, in addition to $|\alpha_{MH}^*| \geq 1$ (Zwick n.d.). Applying the Bonferroni correction here would suggest $p \leq 0.0011$ 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 $\alpha_{MH}^*$ values of 0.14 and 0.39). Both were multiple response questions, one on the topic of mechanics (illustrated in Figure 1a) 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 $\alpha_{MH}^*$ 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, Logan and Gallagher (Hazel, Logan & Gallagher 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. (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, Pollock and Finkelstein (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. (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

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

References

URL: https://doi.org/10.1080/09639284.2011.650447

conceptual understanding of newtonian mechanics: a uk cross-institution comparison, European
URL: http://stacks.iop.org/0143-0807/34/i=2/a=421


Gipps, C. V. & Murphy, Patricia, . (1994). A fair test? : assessment, achievement, and equity,

URL: https://doi.org/10.1080/0950069970190402


