Examining Gender Effects in Different Types of Undergraduate Science Assessment

Joanna Kacprzyk[1*], Martin Parsons[2], Patricia B. Maguire[2] & Gavin S. Stewart[1]

[1]School of Biology and Environmental Science, University College Dublin, Dublin, Ireland
[2]School of Biomolecular and Biomedical Science, University College Dublin, Dublin, Ireland
*corresponding author: joanna.kacprzyk@ucd.ie

ABSTRACT

The optimum assessment structure measures student knowledge accurately and without bias. In this study, the performance of the first-year undergraduate science students from the University College Dublin was evaluated to test the gender equality of the assessment structure in place. Results of male and female students taking three life science modules were analysed, for two academic years, with assessment structure based on a combination of three types of evaluation: continuous assessment and multiple choice questions (MCQ) exam scored with/without negative marking. We found no significant gender effect associated with performance in continuous assessment, or MCQ exams scored without negative marking. However, a significant bias against females was consistently observed for the same cohort of students in the MCQ exams with negative marking of 0.25 points. This bias was at least partially linked to a gender difference in willingness to guess and preliminary data suggest that it disappears after removal of negative marking from the MCQ exams. Our results support the view of a diverse assessment structure being fairer to the students. Moreover, caution is advised while using negative marking, and regular reviews of assessment strategy should be implemented by higher education institutions to ensure gender-bias free evaluation of students' performance.

Keywords: gender-bias; undergraduate science assessment, negative marking

Introduction

The optimum assessment strategy should measure student knowledge correctly and without bias. At the same time, large class sizes, especially in many first-year undergraduate modules, require time efficient methods of assessment. A mixed assessment structure, combining different forms of examination and coursework, is becoming increasingly popular and

dovetails with students' preferences, as they tend to favour coursework based assessment or a mixture of coursework and examinations to the examinations alone (Chamorro-Premuzic et al. 2005; Furnham et al. 2008; reviewed by Richardson [2015]). Using diverse assessment modes is also considered generally fairer to the students (Race and Brown 1998). However, design and implementation of the assessment structure in the undergraduate teaching is rarely followed by the systematic monitoring and quantitative analyses estimating if the created assessment system is indeed equitable, and demonstrates no bias associated with gender or other factors. The literature reports on the gender bias associated with different types of student evaluation are often conflicting, and therefore not always applicable to policy making by learning institutions. This lack of consensus in the literature may be linked to differences between student populations (e.g. in studied subject, level of education or geographical area). In this study we present an empirical approach to evaluation and improvement of current teaching practices. We suggest that staff involved in teaching individual undergraduate programmes can readily monitor for significant gender bias within their assessment structures. We also demonstrate that this can achieved using data generated during routine grading, and with no disruption to student assessment itself. Herein, we analysed the presence of potential gender bias in three stage 1 science undergraduate modules at University College Dublin, each employing the mixed assessment structure. In particular, we focused on the mode of student evaluation. To the best of our knowledge such analysis has not been performed previously in the Republic of Ireland. This study used quantitative data on students' performance over 2 consecutive years that had been collected during routine course assessment. Out of three types of evaluation investigated (continuous assessment and MCQ exam with/without negative marking), only the exams scored with negative marking were associated with significant bias against females. Moreover, our preliminary data suggest that abolishing the negative marking also removed the gender difference in MCQ exam

performance. This demonstrates that regular reviews of the employed assessment strategy can contribute to merit-based, equitable assessment free of gender bias. Our intention is to use this proof-of-concept study to encourage similar local audits of gender equality of the undergraduate student assessment in other Irish Higher Education Institutions to promote teaching excellence and the enhancement of student learning.

Assessment Strategy and Gender Bias: Coursework vs. Examination

Examination and coursework both have their strengths and weaknesses. Unseen examinations eliminate plagiarism and can be easier to mark, especially when automated MCQ exams are considered. However, they also encourage rote learning, provide one-shot only of students' capabilities, and disadvantage students prone to acute stress. Coursework, on the other hand, provides more reliable estimate of student's capabilities, stimulates student learning, especially if feedback is regularly provided and can improve their time management skills. The cons of continuous assessment are increased workload for teaching staff and favouring students able to withstand chronic stress (Brown 2001, 20). Using mixed assessment strategy, within the course, and within a degree programme, is therefore often desirable.

In education there is a general trend of increased emphasis on continuous assessment as opposed to traditional approach based mainly on the high stake terminal examination. This includes the Irish educational system, where the reform of the Junior Cycle (covering the first three years of post-primary education) being currently introduced lists enhanced school-based continuous assessment as an important element promoting student learning and skills development (MacPhail et al. 2018). Likewise, recognising the role of continuous assessment has been postulated for the Leaving Certificate (finalising the Senior Cycle of post-primary education) during the Oireachtas Education Committee meeting in March 2018, by parents, students and universities, despite certain concerns about transparency and objectivity raised by the teachers' representatives. Ensuring that all learners have opportunities to succeed and

reducing student stress are often listed as arguments favouring assessment structure based on combination of coursework and final exams. Indeed, literature links overreliance on high stakes testing with educational inequality, as its negative effect on low economic class and minority students is reported (Au 2008). Importantly, a combination of exams and coursework assessment was described to accompany lower stress levels in Irish Students taking Leaving Certificate Applied (Banks and Smyth 2015). The merits of mixed assessment structure are also acknowledged by the Irish universities. For example, University College Dublin encourages the use of a variety of assessment methods to allow students to demonstrate different types of learning. However, while using diverse methods of assessment is accepted to create fairer assessment structure (Race and Brown 1998), the potential biases associated with each assessment component should also be taken into consideration.

In this analysis we investigated the presence of gender bias in different types of assessment. Anecdotally, female students are favoured by continuous assessment, while the examinations favour males; which is often explained by the average personality predispositions of each gender. For example, such assumptions were made to explain gender achievement patterns in A-level secondary school leaving qualification in UK. Prof. Alan Smithers (Director of the Centre for Education and Employment Research at Buckingham University) told that "Boys tend to do better in exams" and "Girls apply themselves to coursework and work more consistently throughout the year" (Garner 2010). Similarly, in Ireland, Dr Anne Looney, director of the National Council for Curriculum and Assessment suggested that girls' superior performance in the Leaving Certificate was due to "the diligent approach of female students to their studies and exam preparation", but explained that boys do better in multiple choice-style papers because "they take a stab at answers, while girls agonise about which box to tick" (K. Donnelly 2014). However, the literature reports examining these common perceptions tend to vary. For example, coursework favouring

females was suggested in secondary education (reviewed by Elwood 1999) and a similar trend was suggested also in higher education (Chapman 1996; Mallier, Morwood, and Old 1990; McNabb, Pal, and Sloane 2002). In contrast, Elwood (2005) argues that girls appear to do well both in examinations and in coursework assignments and that although coursework has some influence, it is unlikely to be the only cause for the gender gap observed among secondary students in UK. Moreover, Woodfield, Earl-Novell, and Solomon (2005) found little evidence to support the coursework female advantage among UK undergraduate students, showing that female undergraduates outperform male undergraduates, but not as a result of their performance with respect to either mode of assessment in particular. More recent study from the University of Edinburgh indicated that while examination scores show no distinct gender trends, female students show consistently higher coursework scores compared to males across physics, chemistry and biology first year courses (R. C. A. Donnelly 2014). This is in contrast to a study showing that female students underperform on exams compared to their male counterparts in introductory biology courses at the University of Minnesota (Ballen, Salehi, and Cotner 2017). This distinct lack of consensus in the literature suggests the high complexity of the topic, with the gender effect depending on multiple factors, like subject, level of education, geographical area, and potentially, dynamically changing over time. Considering the mixed literature reports on potential gender bias associated with different modes of assessment, we postulate that regular reviews of assessment structure should be performed in educational institutions in order to ensure merit-based, equitable assessment.

Are Multiple Choice Question Examinations Biased Against Females?

The mode of the examination itself can also be biased against particular gender or personality traits. The MCQ exams are a popular assessment method, especially in case of large classes, as they allow examiners to test a wide range of content during a short time and can be scored

quickly either electronically or by hand. They also increase the scoring objectivity by removing the scorer bias. However, the MCQ exam format is frequently criticised for testing the students' knowledge only superficially and several studies suggested that is introduces gender bias in favour of male students. For example, males were demonstrated to perform better than females in economics courses, especially when MCQ exam is the method of evaluation (Ferber, Birnbaum, and Green 1983; Krieg and Uyar 2001; Lumsden and Scott 1987; van Walbeek 2004), but this difference is reduced (Ferber, Birnbaum, and Green 1983; van Walbeek 2004), or even reversed, when essays are used as the test format (Lumsden and Scott 1987). A similar effect was recently observed in accounting examinations, where although women outperformed their male counterparts both in MCQ and constructed-response (e.g. short answer, essay, diagram) formats, their superior performance was diminished in the multiple-choice compared to constructed-response questions (Arthur and Everaert 2012). In contrast, female mathematics students had the same slight advantage over male students in both multiple-choice and constructed-response questions (Wester and Henriksson 2000).

Another concern linked with using the MCQ format is that it offers unprepared students a chance to get credit for content they do not know by simply guessing the answers. A popular strategy to improve ability of MCQ format to distinguish between the weakest and strongest students' performance, is negative marking (Brady 2005; van Walbeek 2004). Students are penalized for incorrect answers and as a consequence, discouraged from guessing, which is expected to increase test reliability and validity, making the test score a truer reflection of a student's ability (Kurz 1999). Nonetheless, the negatively marked MCQ format is not without disadvantages. For example, negative marking is often thought to be biased against higher anxiety students, although the anxiety level seems to be in fact only slightly correlated with the results of negatively marked examinations (Pamphlett and Farnill 1995). Another criticism is that negative marking may disadvantage risk-averse students,

whose confidence levels would discourage them from attempting all questions even if they have at least partial understanding of the topic (Bond et al. 2013). On the other hand, highly gambling prone students may also be disadvantaged by negative marking due to lowered exam scores resulting from wild guessing at the answers. As a result, both groups of students may feel they are penalized for the wrong reason (Bond et al. 2013). Finally, there is certain evidence that using negative marking as a correction for guessing in MCQ exams may introduce gender bias. For example, a type of MCQs, True-False-Abstain exam format was linked with significant gender bias against females in medical education, which is commonly explained by greater risk taking behaviour in males (Kelly and Dennick 2009). Similarly, when penalty for wrong answers was applied, women skipped more questions and received lower SAT II history test scores than men with the same knowledge of the material, which could be partially explained by differences in their risk preferences (Baldiga 2014). A large study of 1947 MCQ negatively marked exams among economics students in Spain also show that women consistently answer less questions than man, and although differences in finals scores are small, penalisation of wrong answers may lead to effective discrimination against women due to their higher risk aversion (Marín and Rosa-García 2011). However, other studies show a lack of gender bias in negatively marked MCQ performance of medical undergraduates (Ricketts, Brice, and Coombes 2010), and life science stage 1 and stage 2 students (Bond et al. 2013). Therefore, the literature evidence for gender bias associated with both MCQ format itself and the negative marking is not consistent and often conflicting, which poses a difficulty in designing a fair assessment strategy.

Gender Profile of Life Science Undergraduate Students

The field of biology is particularly interesting from the point of view of gender equality in the third level education. Unlike in other STEMM (science, technology, engineering, and maths) disciplines, women numerically dominate males accounting for more than 60% of

undergraduate biology majors and approximately half of all graduate students in the biosciences (Amelink 2009; Luckenbill-Edds 2002). Ireland follows the same demographics, where 59% of 2016/2017 new entrants and 60% of 3rd level 2016 graduates in biology and related sciences were women (HEA, 2016, 2017). However, while the number of females studying biology is consistently higher than the number of men, the reports on their performance differ. For example, females were reported to underperform on introductory biology exams compared with males with similar overall college grade point averages and be less active in whole-class discussions (Eddy, Brownell, and Wenderoth 2014). In contrast, other study observed no difference in academic achievement between males and females in introductory biology courses (Lauer et al. 2013). This suggests that generalizations about the gender equality in life science should not be made, and that the parity of educational setting should be evaluated on the case by case basis.

Methods

Data Source

The Stage 1 life sciences modules analysed in this study were BIOL10110 (Cell Biology and Genetics), BIOL10140 (Life on Earth) and BMOL10030 (Biomedical Sciences). Each of these modules employ a mixed assessment structure, based on a combination of continuous assessment and MCQ exams. Students' performance in the academic years of 2014/15 and 2015/16 was analysed. The assessment structure and weighting of individual assessment components is presented in Table 1. Exam and online homework results are collected in the form of scores (%). Graded components (final grade and laboratory practical grade) were converted to % scores using calculation points from the standard component grade scale that is used for all UCD modules (Table S1, supplementary material). A cohort of students taking all 3 modules, and who had attempted all assessment components was selected. In year 2014/15, the results from a total of 188 students were analysed (121 females and 67 males).

In year 2015/16, the results from a total of 151 students were analysed (97 females and 54 males). These representative datasets contained the results of 46 – 83% of all students taking the individual modules in respective academic years and followed the same demographics (typical for biological sciences) of over 60% of students being female. Moreover, they showed the same trends of the relative female-male performance in the individual components of assessment within each module as the whole class data. Analysing these datasets increased the validity of our comparisons of gender related differences in performance associated with differently scored MCQ exams (with negative marking: BIOL10110, BIOL10140; without negative marking: BMOL10030 module), as we were comparing results obtained for each module and module component by exactly the same group of students. Students' results from both years are presented in table S2 A (Supplementary material). The numbers of correct, incorrect and unanswered questions were also recorded for the negatively marked MCQ exams in BIOL10140 and BIOL10110 modules (Table S3 A, Supplementary Material). We have also analysed the students' performance in MCQ exams during the first year (2017/18) of the negative marking for BIOL10110 and BIOL10140 being abolished from the assessment of these modules, which was a decision made by the School of Biology and Environmental Science in June 2017 based on our analysis of 2014/15 and 2015/16 data.

Data Analyses

We followed a bivariate approach to test the hypothesis under examination (i.e. potential gender effect associated with different types of assessment) using data generated during routine grading. In our analyses we limited the student gender to a female/male binary, as these data are currently readily available in student records database. However, we do acknowledge the presence of the gender identities or expressions outside the gender binary, which may in the future be recognised by the UCD's gender policy (Hardesty 2017), but were

not included in this study. We assessed students' performance for each component of each module (including final module grade) by using female-to-male ratio as the indicator of difference between genders (female-to-male ratio = mean female score/mean male score). Female-to-male ratio is a well-established, intuitive and easily understood tool for measuring the scale of potential gender gap, employed for example by the Global Gender Gap Report published by the World Economic Forum (Schwab et al. 2016). Using the relative comparisons, like female-to-male ratio, is not advised in case of very low absolute values where even a small difference can produce a large relative change. In this current study, however, the female-to-male ratio readily shows the relative differences in performance between genders associated with different types of assessment, irrespective of the absolute score values. The mean female-to-male ratios for two analysed academic years (2014/15 and 2015/16) were considered two independent experimental repeats for each assessment component of each module (see table S2 B, Supplementary Material). The significance of the difference between genders was assessed by one sample t-test (female-to-male ratio equal 1, indicating no gender difference, was set as t-test value). The female-to-male ratio for the final grade was also analysed for each module. Similarly, the female-to-male ratio was used to analyse the potential gender gap in percentage of correct answers of all answers given for BIOL10110 and BIOL10140 (see Table S3 B, Supplementary Material). We also tested the pattern of passes (abstain from answer) for the gender effect in the negatively marked MCQ. This was achieved by first recording the percentage distributions of students falling in 3 levels of arbitrary selected pass categories (≤5% 'low' pass, 5-20% 'medium' pass, ≥20% 'high' pass) for males and females (Table S4, Supplementary Material). Subsequently, the female-to-male ratio (% of females/ % of males) was calculated for each pass category, for each academic year (Table S4, Supplementary Material) and the one-way ANOVA, plus Tukey-HSD post-hoc tests, were performed to test for the difference in female-to-male ratio

between the pass categories. Finally, the preliminary data showing students' performance in the MCQ exams for BIOL10110 and BIOL10140 in the first year (2017/18) after the negative marking was removed from the assessment of these modules were analysed by independent samples Mann-Whitney U test. All statistical analyses were performed using SPSS software.

Ethics Statement

The aim of this study was to evaluate and hence suggest potential improvements to current teaching practices, with particular focus on different modes of assessment in place. From the ethical point of view, students could not be separated into two different test groups (e.g. to directly compare MCQ exams with and without negative marking in the same module, in the same year). We therefore had to devise a strategy to compare the cohort of students who were taking different biology modules employing different assessment strategies. The required data were collected as part of routine course assessment and all information was analysed in anonymous form. Data were anonymised by removing student identifiers (name and student number). Ethical approval for this study was received from the University College Dublin Research Ethics Committee (reference number LS-E-15-102-Stewart).

Results

Gender bias and assessment type

In order to address the issue of potential gender bias associated with assessment type (continuous vs exam) and exam mode (MCQ with and without negative marking) we analysed the female-to-male ratio for different assessment components of three stage 1 science modules (Figure 1). The ratios for 2 academic years analysed (2014/15 and 2015/16) considered two independent experimental repeats. Female-to-male ratio equal to 1 indicates no difference in performance of both genders, ratio significantly higher than 1 indicates females outperforming males, while lower than 1 males outperforming females. For all 3

modules analysed, there was no significant gender effect (t-test, see Table S5 in Supplementary Material for all P values and 95% confidence intervals, CI) on continuous assessment score (homework submitted through e-learning platform and practical laboratory reports) (Figure 1 A-C). However, there was a significant gender-dependent effect (t-test: BIOL10140, P=0.038, CI -0.2232, -0.0305; BIOL10110, P=0.031, CI -0.1555, -0.0363), with female-to-male ratio <1, for MCQ exams with negative marking of -0.25 (Figure 1 A, B). Crucially, this difference in gender performance was not observed for the same cohorts of students in the MCQ exams without negative marking (t-test: final MCQ, P=0.965, CI -0.3648, 0.3680; mid-term MCQ, P=0.576, CI -0.0863, 0.0977) which were the part of BMOL10030 assessment (Figure 1C), suggesting that the negative marking may be biased against females. The final grades for modules employing negatively marked MCQs were on average lower for female students (female-to-male ratio <1), but the differences observed were not statistically significant at the 5% level (BIOL10140, P=0.051, CI -0.0986, 0.0009; BIOL10110, P=0.089, CI -0.1189, 0.0335).

[Figure 1 near here]

These results led to elimination of the negative marking from the undergraduate assessment in the School of Biology and Environmental Science starting from the academic year 2017/18. The preliminary data from the first year when MCQ exams for BIOL10140 and BIOL10110 were run with no negative marking indicate an increase of the female-to-male ratio for the final exam to 1.01 and 1.03 respectively. Females (n=119) got the mean score of 63.3% in the BIOL10140 final MCQ, compared to 62.8% achieved by males (n=50), which was not statistically different (independent samples Mann-Whitney U test, P=0.920). Similarly, there was no significant difference between mean BIOL10110 MCQ scores: females (n=226) achieved 67.26% and males (n=89) scored 65.23% (independent samples Mann-Whitney U test, P=0.318).

MCQ with negative marking: higher risk aversion in females?

As previously indicated, females achieved lower scores (female-to-male ratio <1) in the negatively marked MCQ exams. However, there was no evidence of statistically significant gender effect when the percentage of correct answers of all answers given was considered: the mean female-to-male ratio (average of two years when the negative marking was applied) was 0.95 (one sample t-test value=1, P=0.075, CI -0.1279, 0.0261) for BIOL10140 and 0.97 (one sample t-test value=1, P=0.079, CI -0.0708, 0.0162) for BIOL10110. To further explore the observed difference between genders in the scores of negatively marked MCQ exams, we tested if there was a gender specific pattern of passes (students abstaining from giving an answer). Students were divided into following categories: low passers (leaving 5% or less questions blank), medium passers (from 5 to 20 % of questions blank) and high passers (20 % or more questions blank) (Supplementary material, Table S4). Female-to-male ratios were analysed for percentage distributions of female and male students falling into these pass categories (ANOVA, 3 levels), followed by Tukey-HSD post-hoc analysis (Figure 2 A, B). For both modules, a similar trend was observed, with female-to-male ratio significantly higher in high pass category compared to low pass category.

[Figure 2 near here]

Discussion and conclusions

Development of reliable and equitable assessment strategy is often a challenge in the case of stage 1 life science undergraduate modules, which cover a broad range of material and have large class sizes. Assessment is described as an ethical activity (Milligan 1996) and carries a great deal of responsibility (Jarvis 1985; Rowntree 1992). Therefore, the design of assessment structure accurately measuring knowledge without bias against particular groups of students is vital, and should be confirmed by regularly updated, detailed research. The

focus of this study was the empirical investigation of the potential gender bias in different types of the stage 1 science undergraduate assessment system at University College Dublin. Such analysis had not been previously performed, therefore we want to use this proof-of-concept study to create impetus for similar local initiatives in Higher Education Institutions in Republic of Ireland and beyond. In this study a simple bivariate analysis of data generated during routine grading identified the gender effect associated only with the negatively marked MCQ exams. This, in turn, lead to elimination of gender bias from our assessment strategy. Ensuring undergraduate assessment system free of gender bias, which in the long term may contribute to equal access to bursary & scholarship funding, as well as further study opportunities, aligns well with the aims of Athena SWAN. The Athena SWAN Charter programme is aimed at advancing the careers of women in science, technology, engineering, maths and medicine (STEMM) employment in higher education and research. Established in 2005 in UK, Athena SWAN Charter was brought to Ireland in early 2015, and UCD received Athena SWAN Bronze Institutional Award in March 2017. This illustrates a strong incentive to promote gender equality in all aspects of STEMM education and careers in Ireland, which should create a driving force for local audits of gender equality.

Despite the anecdotal evidence and certain claims from literature, this study found no evidence that continuous assessment favours females. The female-to-male ratios were not significantly deviated from 1 for both laboratory report grades and online homework assignments scores. Similarly, there was no significant effect of gender on the MCQ exam scores where no negative marking was applied (BMOL10030 module). As previously highlighted, the only assessment component showing statistically significant bias against female students was MCQ exam with negative marking, with female students achieving only 87.3±0.8 (BIOL10140 module) and 90.4±0.5% (BIOL10110 module) of average male student score. However, due to mixed assessment structure in place for these modules, there

was only a small and not significant difference between overall female and male performance (i.e. final grade achieved). This is in line with study by Lauer et al. (2013) who reported no gender dependent difference in academic achievement in introductory biology courses. It also supports the view that the greater the diversity in the methods of assessment, the fairer the assessment is to the students (Race and Brown 1998) and that overreliance on one assessment mode could be linked with risk of favouring particular students, resulting in one-dimensional or limited evaluation of student performance (Brady 2005).

We clearly acknowledge that the analysis presented herein does not account for the number of variables, which may influence the gender differences in student achievement, such as ethnicity, age or previous attainment. A larger multivariate study involving higher student numbers is required to fully explore an array of possible explanatory factors. Nevertheless, the bias against female students observed in scores of negatively marked MCQ exams (despite the relatively moderate penalty of -0.25 point/incorrect answer) casts doubts over the validity of this exam format and suggests that it should be treated with caution, especially considering that the same cohort of students showed no difference in their MCQ scores where no negative marking was applied. Based on this analysis, the negative marking was removed from the MCQ exams in the School of Biology and Environmental Science at UCD, starting from the academic year 2017/18. Preliminary data from the BIOL10110 and BIOL10140 modules suggest that the gender difference in the MCQ exam achievement was eliminated by removing the negative marking, further supporting the view of this scoring method being biased against female students. The gender bias against females in negatively marked examination formats has been reported before (Kelly and Dennick 2009), while other studies denied it (Bond et al. 2013; Ricketts, Brice, and Coombes 2010). It may be therefore argued that the gender bias (or lack thereof) associated with negatively marked MCQ, or other type of assessment, depends on the subject, cultural factors, stage of education, and

many other variables. Consequently, achieving and maintaining gender equality should involve regular reviews of assessment strategies carried out locally by higher education institutions. The case study presented herein demonstrates that simple local audits of gender equality can lead to informed decisions aimed at improvements of assessment structure in use. This practice is already encouraged in education-leading universities – such as Ghent University (45th Life Science University in the World, 17th in Europe), which dismissed negative marking in multiple choice assessment starting from 2014/15. Their decision was based on literature, but also on research of uGhent data and probability analysis, and made to eliminate of bias linked with personality features (e.g. tendency to guess) (Ghent University 2013).

We attempted to explain the observed difference in negatively marked MCQ scores by analysing the gender specific patterns of passes (abstaining from answer). The female-to-male ratio was significantly higher in the high pass category (proportion of students giving 20% or more blank answers) compared to low pass category (proportion of students giving 5% or less blank answers). These data suggest that, despite the principles of negatively marked exams being repeatedly explained during the relevant modules, female students appeared more likely to leave questions unanswered to avoid penalty. Baldiga (2014) proposes several explanations for this unwillingness to guess typical for women taking MCQ exams. One of them is that women skip more questions than men simply because they know less about the material. This does not seem to be the case in our study, as removing the negative marking eliminated the gender difference in students' performance, and also due to the lack of gender gap in MCQ exams without negative marking in BMOL10030 module. Other explanations include higher risk-aversion, lower confidence or differences in responses to high pressure in women (Baldiga 2014). Without collecting additional data, it is not possible to hypothesize which of them contributed to results observed in this study. However,

the most widely accepted explanation in literature is that the gender-specific risk aversion is linked to female underperformance when assessment penalizes incorrect answers (Baldiga 2014; Burns, Halliday, and Keswell 2012; Davies, Mangan, and Telhaj 2005; Marín and Rosa-García 2011).

In conclusion, presented data contribute to the discussion about gender differences across modes of assessment. We demonstrate that a simple analysis of potential gender bias can be easily performed on the components of routine assessment and despite the limitations of bivariate approach, it may offer a valuable tool for achieving gender equality in higher education. Results suggest a bias against females linked with negative marking and reinforce the validity of mixed assessment structure, as being most fair to the students. By this study we want to encourage similar activities aimed at monitoring and elimination of potential gender bias in the undergraduate science teaching, which can in the long term contribute to improved persistence of female students in STEMM disciplines. Such incentives align well with the aims of Athena SWAN Charter and demonstrate commitment to creating truly equitable learning environment.

Disclosure statement

We declare no conflict of interests

Notes on contributors

Joanna Kacprzyk is an assistant professor in the School of Biology and Environmental Science at University College Dublin. She is an active member of the undergraduate science teaching team. Her primary research interests are cellular stress responses for programmed cell death and survival.

Martin Parsons is a postgraduate researcher in the School of Biomolecular and Biomedical Science. He is a member of the undergraduate biochemistry practical teaching team and his research is focused on providing a simple diagnostic test for multiple sclerosis.

Patricia B. Maguire is an associate professor in Biochemistry in the School of Biomolecular and Biomedical Science. A major focus of her undergraduate teaching has been stage 1 restructuring while her research looks at the changing messages that platelets carry in inflammatory diseases.

Gavin S. Stewart is an assistant professor in the School of Biology and Environmental Science at University College Dublin. He is the departmental Head of Teaching and Learning, while the primary interest of his research group are the physiological roles of facilitative urea transporters.

Funding

This work was supported by P.B.M. and M.P. were supported by research fellowship award by the School of Biomolecular and Biomedical Science, UCD.; J.K. was partially funded by ERC-2012-StG311000 grant awarded to Prof. Emma C. Teeling.

References

Amelink, Catherine. 2009. "Literature Overview: Gender Differences in Science Achievement, SWE-AWE CASEE Overviews." *University Park, PA: Assessing Women and Men in Engineering Project*

Arthur, Neal, and Patricia Everaert. 2012. "Gender and Performance in Accounting Examinations: Exploring the Impact of Examination Format." *Accounting Education* 21 (5):471-487. doi: 10.1080/09639284.2011.650447

Au, Wayne W. 2008. "Devising inequality: a Bernsteinian analysis of high-stakes testing and social reproduction in education." *British Journal of Sociology of Education* 29 (6):639-651. doi: 10.1080/01425690802423312.

Baldiga, Katherine. 2014. "Gender Differences in Willingness to Guess." *Management Science* 60 (2):434-448. doi: 10.1287/mnsc.2013.1776

Ballen, Cissy J., Shima Salehi, and Sehoya Cotner. 2017. "Exams Disadvantage Women in Introductory Biology." *PLOS ONE* 12 (10):e0186419. doi: 10.1371/journal.pone.0186419k.

Banks, Joanne, and Emer Smyth. 2015. "'Your whole life depends on it': academic stress and high-stakes testing in Ireland." *Journal of Youth Studies* 18 (5):598-616. doi: 10.1080/13676261.2014.992317.

Bond, A. Elizabeth, Owen Bodger, David O. F. Skibinski, D. Hugh Jones, Colin J. Restall, Edward Dudley, and Geertje van Keulen. 2013. "Negatively-Marked MCQ Assessments That Reward Partial Knowledge Do Not Introduce Gender Bias Yet Increase Student Performance and Satisfaction and Reduce Anxiety." *PloS One* 8 (2):e55956. doi: 10.1371/journal.pone.0055956.

Brady, Anne-Marie. 2005. "Assessment of Learning with Multiple-Choice Questions." *Nurse Education in Practice* 5 (4):238-42. doi: http://dx.doi.org/10.1016/j.nepr.2004.12.005.

Brown, George. 2001. Assessment: A guide for lecturers. Vol. 3: LTSN LTSN Generic Centre, York

Burns, Justine, Simon Halliday, and Malcolm Keswell. 2012. "Gender and Risk Taking in the Classroom." *Southern Africa Labour and Development Research Unit, University of Cape Town* Working Paper 87.Chamorro-Premuzic, Thomas, Adrian Furnham, Georgia Dissou, and PatrickHeaven. 2005. "Personality and Preference for Academic Assessment: A study with Australian University Students." *Learning and Individual Differences* 15 (4):247-56. doi: http://dx.doi.org/10.1016/j.lindif.2005.02.002.

Chapman, Keith. 1996. "An Analysis of Degree Results in Geography by Gender." *Assessment & Evaluation in Higher Education* 21 (4):293-311. doi: 10.1080/0260293960210401.

Davies, Peter, Jean Mangan, and Shqiponja Telhaj. 2005. "Bold, Reckless and Adaptable? Explaining Gender Differences in Economic Thinking and Attitudes." *British Educational Research Journal* 31 (1):29-48.

Donnelly, Katherine. 2014. "Girls Take the Lead over Boys in Leaving Cert Exams" Independent, August 15. https://www.independent.ie/irish-news/education/girls-take-the-lead-over-boys-in-leaving-cert-exams-30510201.html

Donnelly, Robyn C. A.. 2014. "Gender Differences in Undergraduate Students' Performance, Perception and Participation in Physics." PhD diss., University of Edinborough.

Eddy, Sarah L., Sara E. Brownell, and Mary P. Wenderoth. 2014. "Gender Gaps in Achievement and Participation in Multiple Introductory Biology Classrooms." *CBE Life Sciences Education* 13 (3):478-92. doi: 10.1187/cbe.13-10-0204.

Elwood, Janette. 1999. "Equity Issues in Performance Assessment: The Contribution of Teacher-Assessed Coursework to Gender-Related Differences in Examination Performance." *Educational Research and Evaluation* 5 (4):321-44. doi: 10.1076/edre.5.4.321.6937.

Elwood, Janette. 2005. "Gender and achievement: what have exams got to do with it?" *Oxford Review of Education* 31 (3):373-93. doi: 10.1080/03054980500222031.

Ferber, Marianne A., Bonnie G. Birnbaum, and Carole A. Green. 1983. "Gender Differences in Economic Knowledge: A Reevaluation of the Evidence." *The Journal of Economic Education* 14 (2):24-37. doi: 10.2307/1182793.

Furnham, Adrian, Andrew Christopher, Jeanette Garwood, and Neil G. Martin. 2008. "Ability, Demography, Learning Style, and Personality Trait Correlates of Student Preference for Assessment Method." *Educational Psychology* 28 (1):15-27. doi: 10.1080/01443410701369138.

Garner, Richard. 2010. "The smarter sex: Does it matter if girls do better than boys?" *Independent*¸ October 20. http://www.independent.co.uk/news/education/schools/the-smarter-sex-does-it-matter-if-girls-do-better-than-boys-2112129.html
Ghent University. 2013. "*No more negative marking in multiple choice questions*." https://www.ugent.be/student/en/class-exam-exchange-intern/class-exam/OEREnglish/multiplechoice.htm#reasonHardesty, Aoife. 2017. "Gender Recognition on the Way for Transgender People in UCD?" *University Observer*¸ February 6. http://www.universityobserver.ie/features/gender-recognition-on-the-way-for-transgender-people-in-ucd/
HEA (Higher Education Authority) 2016. "*All Undergraduate Graduates by Level and Field of Study, 2016*" http://hea.ie/statistics-archive/
HEA (Higher Education Authority) 2017. "*New Entrants by Institution, Gender and Field of Study (ISCED), 2016/17*" http://hea.ie/statistics-archive/
Jarvis, Peter. 1985. *The sociology of adult and continuing education*. Glasgow: Routledge.
Kelly, Shona, and Reg Dennick. 2009. "Evidence of Gender Bias in True-False-Abstain medical examinations." *BMC Medical Education* 9 (1):32. doi: 10.1186/1472-6920-9-32.
Krieg, Randall G., and Bulent Uyar. 2001. "Student Performance in Business and Economics Statistics: Does Exam Structure Matter?" *Journal of Economics and Finance* 25 (2):229-241. doi: 10.1007/bf02744525
Kurz, Terri B. 1999. "A Review of Scoring Algorithms for Multiple Choice Tests." Paper presented at the Annual meeting of the Southwest Educational Research Association, San Anton.
Lauer, Shanda, Jennifer Momsen, Erika Offerdahl, Mila Kryjevskaia, Warren Christensen, and Lisa Montplaisir. 2013. "Stereotyped: Investigating Gender in Introductory Science Courses." *CBE Life Sciences Education* 12 (1):30-8. doi: 10.1187/cbe.12-08-0133
Luckenbill-Edds, Louise. 2002. "The Educational Pipeline for Women in Biology: No Longer Leaking?" *Bioscience* 52 (6):513-21. doi: 10.1641/00063568(2002)052[0513:TEPFWI]2.0.CO;2.
Lumsden, Keith G., and Alex Scott. 1987. "The Economics Student Reexamined: Male-Female Differences in Comprehension." *The Journal of Economic Education* 18 (4):365-375. doi: 10.2307/1182118.
MacPhail, Ann, John Halbert, and Hal O'Neill. 2018. "The Development of Assessment Policy in Ireland: A Story of Junior Cycle Reform." *Assessment in Education: Principles, Policy & Practice*:1-17. doi: 10.1080/0969594X.2018.1441125.
Mallier, Tony, Steven Morwood, and John Old. 1990. "Assessment Methods and Economics Degrees." *Assessment & Evaluation in Higher Education* 15 (1):22-44. doi: 10.1080/0260293900150103.
Marín, Carmen, and Alfonso Rosa-García. 2011. "Gender Bias in Risk Aversion: Evidence from Multiple Choice Exams." Working Paper 39987, MPRA
McNabb, Robert, Sarmistha Pal, and Peter Sloane. 2002. "Gender Differences in Educational Attainment: The Case of University Students in England and Wales." *Economica* 69 (275):481-503.
Milligan, Frank. 1996. "The Use of Criteria-Based Grading Profiles in Formative and Summative Assessment." *Nurse Education Today* 16 (6):413-8. doi: http://dx.doi.org/10.1016/S0260-6917(96)80047-5.
Pamphlett, Roger, and Douglas Farnill. 1995. "Effect of Anxiety on Performance in Multiple Choice Examination." *Medical Education* 29 (4):297-302.
Race, Phil, and Sally Brown. 1998. *The lecturer's toolkit*. Kogan Page London.

Richardson, John T. E. 2015. "Coursework Versus Examinations in End-of-Module Assessment: a Literature Review." *Assessment & Evaluation in Higher Education* 40 (3):439-55. doi: 10.1080/02602938.2014.919628.

Ricketts, Chris, Julie Brice, and Lee Coombes. 2010. "Are Multiple Choice Tests Fair to Medical Students with Specific Learning Disabilities?" *Advances in Health Sciences Education* 15 (2):265-75. doi: 10.1007/s10459-009-9197-8.

Rowntree, Derek. 1992. *Assessing students: how shall we know them?* London: Kogan Page.

Schwab, Klaus, Richard Samans, Saadia Zahidi, Till A. Leopold, Vesselina Ratcheva, Ricardo Hausmann, and Laura D'Andrea Tyson. 2016. "The Global Gender Gap Report 2016." Paper presented at the World Economic Forum.

van Walbeek, Cornea. 2004. "Does Lecture Attendance Matter? Some Observations from a First-Year Economics Course at the University of Cape Town." *South African Journal of Economics* 72 (4):861-883. doi: doi:10.1111/j.1813-6982.2004.tb00137.x.

Wester, Anita, and Widar Henriksson. 2000. "The Interaction between Item Format and Gender Differences in Mathematics Performance Based on TIMSS Data." Studies in Educational Evaluation 26 (1):79-90. doi: https://doi.org/10.1016/S0191-491X(00)00007-9.

Woodfield, Ruth, Sarah Earl-Novell, and Lucy Solomon. 2005. "Gender and Mode of Assessment at University: Should we Assume Female Students are Better Suited to Coursework and Males to Unseen Examinations." *Assessment & Evaluation in Higher Education* 30 (1): 25-50. doi: 10.1080/0260293042003243887.

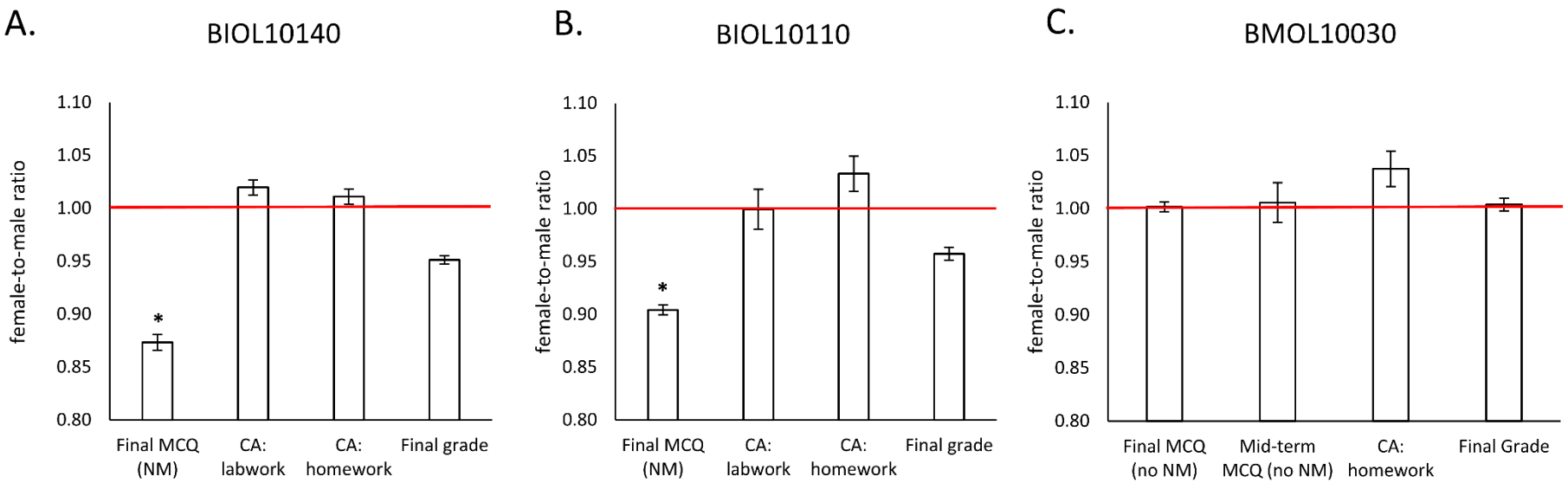

Figure 1. Female-to-male ratio for different assessment types. Female-to-male ratios (means of 2 experimental repeats = 2 academic years, ±SEM) are presented for assessment components of three stage 1 life science modules. The difference between genders was analysed by one sample t-test, with female-to-male ratio equal 1 (horizontal line), indicating no gender difference, set as test value. Asterix (*) indicates p-value <0.05. NM: negative marking, no NM: no negative marking, CA: continuous assessment.

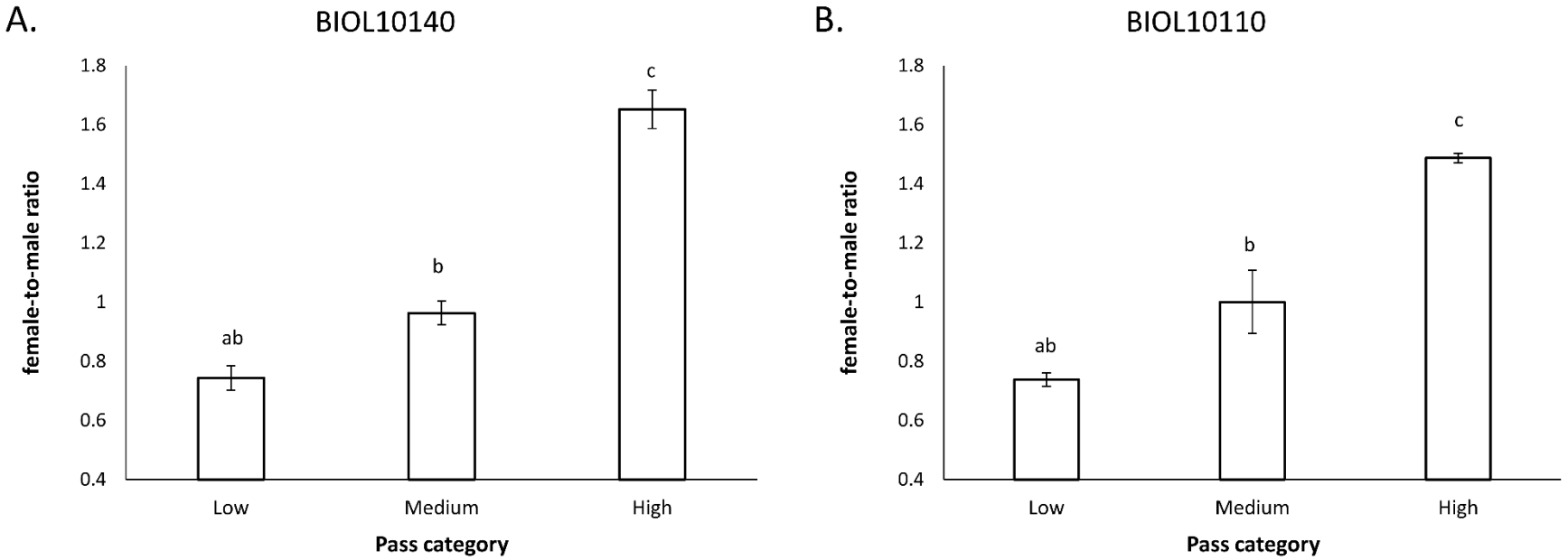

Figure 2. Female-to-male ratio and blank answers in negatively marked MCQ. Female-to-male ratios were calculated for fractions of female and male students falling into low pass (5% or less blank answers), medium pass (5 to 20 % blank answers) and high pass (20% or more blank answers) categories for BIOL10140 (A) and BIOL10110 (B) modules. Means of 2 experimental repeats (2 academic years), ±SEM are presented. Results were analysed by one-way ANOVA (3 pass categories) plus Tukey-HSD post-hoc tests. Means with different letters are significantly different from each other ($p<0.05$).

Table 1. Three life science modules analysed in this study and their assessment structure.

Module	**Semester**	**Assessment Component**	**Weight**
BIOL10140	1	Final MCQ (negative marking - 0.25), 50 questions	50%
		Coursework: laboratory practicals	35%
		Coursework: homework e-learning	15%
BIOL10110	2	Final MCQ (negative marking - 0.25), 60 questions	60%
		Coursework: laboratory practicals	20%
		Coursework: homework e-learning	20%
BMOL10030	2	Final MCQ (no negative marking), 60 questions	70%
		Mid-term MCQ (no negative marking)	20%
		Coursework: homework e-learning	10%

Supplemental table S1. The standard component grade scale that is used for UCD modules.

Grade	Calculation point
A+	78.33
A	75.00
A-	71.67
B+	68.33
B	65.00
B-	61.67
C+	58.33
C	55.00
C-	51.67
D+	48.33
D	45.00
D-	41.67
E+	38.33
E	35.00
E-	31.67
F+	28.33
F (FM)	25.00
F-	21.67
G+	18.33
G	15.00
G-	11.67
NG	0.00

Supplemental table S2. Mean female and male student achievement in three stage 1 science modules used to calculate female-to-male ratios (A) were used to calculate female-to-male ratios (mean female score/mean male score) for respective academic years, treated as two experimental repeats (B).

A)

academic year 2014/15

BIOL10140	Gender	Score (%)	BIOL10110	Gender	Score (%)	BMOL10030	Gender	Score (%)
final MCQ (negative marking)	F	57.4	final MCQ (negative marking)	F	58.8	final MCQ (no negative marking)	F	63.8
	M	65.2		M	65.4		M	65.6
continuous assessment (labwork)	F	70.0	continuous assessment (labwork)	F	69.7	midterm MCQ (no negative marking)	F	72.3
	M	69.1		M	68.4		M	71.4
continuous assessment: homework	F	85.9	continuous assessment: homework	F	89.3	continuous assessment: homework	F	86.5
	M	85.6		M	85.1		M	81.5
final grade	F	64.7	final grade	F	65.9	final grade	F	66.7
	M	68.3		M	68.4		M	67.6

academic year 2015/16

BIOL10140	Gender	Score (%)	BIOL10110	Gender	Score (%)	BMOL10030	Gender	Score (%)
final MCQ (negative marking)	F	51.5	final MCQ (negative marking)	F	54.5	final MCQ (no negative marking)	F	65.9
	M	58.8		M	59.4		M	63.9
continuous assessment (labwork)	F	70.2	continuous assessment (labwork)	F	67.7	midterm MCQ (no negative marking)	F	69.7
	M	68.3		M	68.5		M	69.8
continuous assessment: homework	F	84.5	continuous assessment: homework	F	87.2	continuous assessment: homework	F	95.8
	M	82.0		M	85.7		M	94.5
final grade	F	63.3	final grade	F	63.0	final grade	F	68.8
	M	65.7		M	65.3		M	67.3

B)

Academic year	BIOL10140	female-to-male ratio	BIOL10110	female-to-male ratio	BMOL10030	female-to-male ratio
2014/15	final MCQ (negative marking)	0.88	final MCQ (negative marking)	0.90	final MCQ (no negative marking)	0.97
2015/16		0.87		0.91		1.03
2014/15	continuous assessment (labwork)	1.01	continuous assessment (labwork)	1.02	midterm MCQ (no negative marking)	1.01
2015/16		1.03		0.98		1.00
2014/15	continuous assessment: homework	1.00	continuous assessment: homework	1.05	continuous assessment: homework	1.06
2015/16		1.02		1.02		1.01
2014/15	final grade	0.95	final grade	0.96	final grade	0.99
2015/16		0.96		0.95		1.02

Supplemental table S3. Mean number of correct, incorrect and unanswered (pass) questions given in the negatively marked MCQ exams by female and male students (A) were used to calculate mean % of correct answers of all answers given and female-to-male ratios for % correct answers of all answers given for respective academic years, treated as two experimental repeats (B).

A)

academic year 2014/15

BIOL10140			BIOL10110		
Gender	**Number of answers**		**Gender**	**Number of answers**	
F	correct	31.1	F	correct	38.3
M		34.8	M		42.1
F	incorrect	9.8	F	incorrect	12.0
M		8.9	M		11.4
F	pass	9.1	F	pass	9.7
M		6.3	M		6.5

academic year 2015/16

BIOL10140			BIOL10110		
Gender	**Number of answers**		**Gender**	**Number of answers**	
F	correct	29.2	F	correct	35.8

M		32.1	M		38.5
F	incorrect	13.2	F	incorrect	12.4
M		11.9	M		12.1
F	pass	7.6	F	pass	11.8
M		6.0	M		9.4

B) **BIOL10140** **BIOL10110**

Academic year	Gender	% correct answers of all given	female-to-male ratio	% correct answers of all given	female-to-male ratio
2014/15	F	76.1	0.96	76.2	0.97
	M	79.6		78.6	
2015/16	F	68.9	0.94	74.2	0.98
	M	73.0		76.0	

Supplemental table S4. Percentage distributions of female and male students categorized as low (5% or less blank answers), medium (5 to 20 % blank answers) and high passers (20% or more blank answers) in the negatively marked MCQ exams for BIOL10140 and BIOL10110 modules. Female-to-male ratios (% of females/% of males) were also calculated for each category.

academic year 2014/15

	BIOL10140			BIOL10110		
Pass category	**Female %**	**Male %**	**female-to-male ratio**	**Female %**	**Male %**	**female-to-male ratio**
low pass (<=5%)	29.17	41.51	0.70	22.92	32.08	0.71
medium pass (5-20%)	41.67	41.51	1.00	35.42	39.62	0.89
high pass (>=20%)	29.17	16.98	1.72	41.67	28.30	1.47
Total	100	100		100	100	

academic year 2015/16

	BIOL10140	BIOL10110	

Pass category	Female %	Male %	female-to-male ratio	Female %	Male %	female-to-male ratio
low pass (<=5%)	28.0	37.0	0.76	36.00	49.00	0.73
medium pass (5-20%)	37.2	40.3	0.92	33.06	29.85	1.11
high pass (>=20%)	35.5	22.4	1.59	31.40	20.90	1.50
Total	100	100		100	100	

Supplemental table S5. One sample t-test statistics for investigating the difference between genders in assessment components and final grade of 3 first year undergraduate science modules. Female-to-male ratio equal 1, indicating no gender difference, was set as t-test value. Components significant at 5% level are highlighted in yellow.

BIOL10140	P-value	95% Confidence Interval of the Difference		BIOL10110	P-value	95% Confidence Interval of the Difference		BMOL10030	P-value	95% Confidence Interval of the Difference	
		Lower	Upper			Lower	Upper			Lower	Upper

final MCQ (negative marking)	0.038	-0.2232	-0.0305	final MCQ (negative marking)	0.031	-0.1555	-0.0363	final MCQ (no negative marking)	0.965	-0.3648	0.3680
continuous assessment (labwork)	0.224	-0.0721	0.1114	continuous assessment (labwork)	0.984	-0.2389	0.2379	midterm MCQ (no negative marking)	0.576	-0.0863	0.0977
continuous assessment: homework	0.371	-0.0810	0.1029	continuous assessment: homework	0.297	-0.1792	0.2456	continuous assessment: homework	0.364	-0.2696	0.3448
final grade	0.051	-0.0986	0.0009	final grade	0.089	-0.1189	0.0335	final grade	0.856	-0.2160	0.2239