Gender bias in first-year multiple-choice physics examinations

M. J. Gladys,[1,2] J. E. Furst,[2] J. L. Holdsworth,[2] and P. C. Dastoor[1,2]
[1]*Centre for Organic Electronics, University of Newcastle, Callaghan, NSW 2308, Australia*
[2]*Department of Physics, University of Newcastle, Callaghan, NSW 2308, Australia*

(Received 7 December 2022; accepted 10 July 2023; published 11 August 2023)

The multiple-choice section of the final examination for the first-year Advanced Physics I course at the University of Newcastle, Australia between 2010 and 2018 was investigated for gender bias. A Mantel-Haenszel analysis revealed that approximately 20% of the multiple-choice questions exhibited statistically significant gender bias. A schema for characterizing the multiple-choice questions was proposed and used to analyze the entire question set. Male bias questions showed moderate to large bias and tended to include characteristics related to visualization, though not images. Several questions exhibited a moderate bias in favor of females and were characterized by requiring a numerical calculation involving a simple one-step equation. These results indicate that with continued development, gender bias analysis of physics questions based on a characterization schema may be used as a routine tool for testing for the presence and origin of gender gaps in student performance.

DOI: 10.1103/PhysRevPhysEducRes.19.020109

I. INTRODUCTION

Gender disparity in the participation (going to class) [1], performance (attainment) [2], and outcomes (degrees and careers) [3] of female students in physics is an ongoing concern. Generational changes to student cohorts and society influences over recent decades means that continued study of gender gaps is essential as the outcomes are continually shifting [4]. The disparity in the participation of girls in physics and mathematics is established at an early stage in the Australian education system, with female participation rates decreasing such that by their final school year males outnumber females 3 to 1 in physics and almost 2 to 1 in mathematics [5]. Indeed, changes made to the New South Wales (NSW) secondary education, high school certificate syllabus in 2000 were focused on contextualizing physics, in part to increase female participation [6].

Historically, the disparity in the attainment of females versus males in physics assessments is well documented in the literature. Internationally, female students consistently underperform relative to corresponding male cohorts in undergraduate physics programs [7] as well as in established concept test regimes [8,9].

Disparities in participation and attainment drive differences in the outcomes for female students in physics and other science, technology, engineering, and mathematics (STEM) related subjects, such as information technology, which also display similar disparities in western countries. These differences contribute to the gender inequity in tertiary STEM education and the consequent STEM-based workforce, with twice as many male students aspiring to a STEM related career than females [10,11]. For example, in 2016, women in Australia comprised only 31% of STEM academic and research staff, as well as enduring a 12.6% gap in pay in science positions [12]. In contrast, however, in nonanglophone countries the gender participation gap is, in some cases, reversed [13,14].

In terms of assessment, there is bias in a question "if a factor other than ability (in this case gender) affects the likelihood that a student will answer the question correctly" [15]. Determining whether there is a meaningful difference in student responses due to a bias factor necessarily requires a statistical approach, and there are several statistical methods for quantifying the magnitude and significance of the difference between two groups (see Halpern *et al.* [16] for a comprehensive review of gender studies over the past few decades). On the areas that influence gendered exam performance, studies can be grouped into three main categories: (a) physiological differences, language, and comprehension skills, (b) testing environments, and (c) previous understanding, learning environments, and stereotyping.

A. Physiological differences, language, and comprehension skills

Studies have shown that, males are, on average, more comfortable at spatiomechanical functioning and

visualization while females on average can multitask better and overall have increased reading comprehension [16]. Alternatively, some investigations argue that these differences are not biological but rather must include in part or purely a strong sociocultural component (see Ref. [17], and references therein).

In Australia, girls outperform boys in reading by the equivalent of around one year of schooling [18], which, based on these studies, would suggest that girls should also outperform boys on average in science. Indeed, it is generally agreed that science performance is highly correlated to the degree of a student's reading comprehension [19,20]. Students struggling with reading find difficulties in answering scientific questions due to inadequate comprehension. However, this lack of comprehension is not linked to their scientific understanding which may be equivalent to those designated as "good readers" [21]. Although several studies have demonstrated that grammar can affect exam performance [22,23], the debate continues as to whether word count plays a significant role in correctly answering multiple-choice and short answer exam type questions with opposing study outcomes evident in the literature [24,25].

B. Previous understanding, learning environments, and stereotyping

Wilson *et al.* investigated gender differences in attainment for Australian high school students (selected based on their physics aptitude) undertaking the Australian Science Olympiad Physics Exam over an eight-year period [26]. The study revealed that the majority of multiple-choice questions (MCQs) exhibited at least a small male bias and that the bias was significantly larger for questions with a concrete context or involving visualization or interpreting diagrams. In addition to question bias, the origin of the gender gaps was attributed to females in the cohort having less physics content and procedural knowledge, or the ability to apply that knowledge. Henderson *et al.* investigated gender gaps in the performance of an electricity and magnetism concept survey [27]. Their study revealed that while there was no performance difference in gender in quantitative test questions, qualitative (or concept) problems increasingly showed a gender gap. Moreover, the performance differences were not a result of psychological factors, such as science anxiety or stereotype threat.

There are numerous learning environment and stereotyping experiences that impact on girls and women which may manifest in a decrease in final exam performance [28,29]. For example, experiences of discrimination in the learning environment have a strong negative impact on female persistence in continuing university physics studies [30]. While, in somewhat broad terms, gender conditioning is an important parameter that may generate differences in language interpretation of both quantitative and concept questions its influence may be difficult to distinguish from other contributions detailed previously [31]. A very recent study by Kalender *et al.* [32] revealed that female students in physics related courses were more likely to have a mindset that an innate physics talent was required to do well in physics and that they did not possess this natural ability. In addition, Traxler *et al.* [33] also argue that the intersection of identities plays an important role in understanding the causality and identification of gender bias.

C. Testing environments

Gibson *et al.* investigated the impact of question structure for first-year natural sciences physics graduates from the University of Cambridge [34]. Their study showed that question scaffolding improves performance for all students; however, the average mark improvement favors female students (13.4%, $N = 77$) over male students (9%, $N = 236$) by more than 4%. In contrast, Dawkins *et al.* studied elements of question structure promoting male bias and scaffolding but revealed that the level of scaffolding could not sufficiently explain the gender gap observed [35]. Hedgeland *et al.* investigated whether MCQs are inherently biased and concluded that the use of the MCQ format is not a significant factor in gender gap in assessment [36]. In contrast, a Stanford study found an increased gender gap for examinations dominated by multiple-choice questions rather than open ended questions [37]. More recently, Wilson *et al.* found that changing the way information is presented and adjusting question context could eliminate performance bias in MCQ tests [26]. Based on their study, Salehi *et al.* go even further and suggest moving away from final exams altogether as test anxiety disproportionality impacts females [38].

Thus, the differences in attainment performance between males and females in studying physics may arise from inherent bias, historical male dominance in the content (relevance), differences between intrinsic thought processes, or the way questions are interpreted based on possible male-dominated grammar in exam questions. However, a recent study by Dew *et al.* (analyzing over 10 000 introductory physics examination results over ten years) found no gender bias, suggesting that gendered differences in performance in formal examinations may not be so clearcut and highlighting the need for a detailed examination of the role that individual physics topics play in determining gender attainment gaps [39].

Investigations on the impact of the Covid-19 pandemic have revealed contrasting outcomes in terms of gender gaps in educational performance. A distance education study in economics [40] found that while the gender gap in performance reduced during the lockdown period, post lockdown found that the gap exacerbated even compared to prepandemic outcomes. Several other studies have found either no effect or a positive effect on women's attainment during the pandemic period [41–43]. In terms of test

anxiety, outlined earlier, removal of testing during the pandemic may have reduced test anxiety; however, there appears to be an increase in anxiety from the forced online classroom environment that was hurriedly set up to accommodate learning through the pandemic [44].

Here, our focus is on identifying and characterizing gender bias MCQs that influence the performance (or attainment) of undergraduate students in first-year physics final examinations. This study investigates the outcomes for individual MCQs over an eight-year period of first-year physics final exams at the University of Newcastle, focusing on binary attainment gaps across different physics topics.

II. METHODOLOGY

In this study gender was defined based on whether, on entry to the University of Newcastle, students associated with the two categories male or female. No other descriptors were used on application forms during the years for this investigation. The Advanced Physics I course Phys1210 predominately includes students from science, electrical engineering, and teaching degrees with approximately 200 students per year. The course covers the following topics in order: particle physics, cosmology, mechanics, thermal physics, nuclear physics, oscillations, and, finally, waves. From 2010 to 2014 the mechanics section contained a subsection on special relativity, which was then replaced by advanced rotational motion. In the final two years of the study, the electricity section replaced nuclear physics to assist the electrical engineers participating in the course. Over the years studied approximately 11% of students were female. The final exam comprised 40 multiple-choice questions worth one mark each (total 40 marks), and a short answer section, worth in total 60 marks. The same MCQs were used throughout the period of the study. There were enough students of each gender to allow the use of robust statistical tests.

A. Statistical analysis

A two-sided unpaired t-test showed no significant difference in the mean performance in the course as a whole ($t = 0.25$, $p = 0.8006$, $N = 1414$). However, there was a significant difference ($t = 2.52$, $p = 0.0126$, $N = 1415$) in the multiple-choice section of the final exam with males doing better than females in this section. The MCQ section of the final exam only makes up 20% of the overall assessment for the course, and therefore the overall result is dictated by the statistics from the remaining 80% of course assessment.

One potential issue limiting the interpretation of any statistical test to analyze these types of data is that there may be significant differences in ability within the cohort of students. These differences may not be distributed evenly between the two groups, which could affect the result.

TABLE I. 2×2 contingency table for multiple-choice responses. For the i th stratum a_i is the number of correct male responses, b_i is the number of correct female responses, c_i is the number of incorrect male responses, d_i is the number of incorrect female responses, and N is the total number of responses, $n_1 + n_2$.

Correct response	Male	Female	Total
Yes	a	b	m_1
No	c	d	m_2
Total	n_1	n_2	N

The use of techniques [such as the Mantel-Haenszel (MH) test [45,46]] which stratify the cohort and compare students of equal ability provide a reliable test of bias, as well as accommodating smaller (∼100) overall sample sizes [15,45]. In this study we use the MH test with five strata based on overall score in the multiple-choice section of the final exam. In other words, students were grouped into total exam scores that were roughly the same and questions were then compared to identify whether men and women answered them correctly at the same or different rates.

The MH test uses 2×2 contingency tables based on stratified ability. Table I shows the structure of an individual contingency table. There is a contingency table for each stratum within a question.

The weighted average of the odds ratios from each stratum is given by the Mantel-Haenszel odds ratio:

$$\alpha_{\mathrm{MH}} = \left(\frac{\sum_{i=1}^{K} \frac{a_i d_i}{N_i}}{\sum_{i=1}^{K} \frac{c_i b_i}{N_i}} \right). \tag{1}$$

And the log transformed Mantel-Haenszel Odds ratio is

$$\alpha^*_{\mathrm{MH}} = -2.35 \ln(\alpha_{\mathrm{MH}}). \tag{2}$$

With this transformation the sign and magnitude of α^*_{MH} signify the direction and strength of bias within a question. Therefore $\alpha^*_{\mathrm{MH}} = 0$ indicates no difference between genders, a negative value reveals a bias toward males, and a positive value indicates a bias toward females.

To secure an addition level of significance the α^*_{MH} is tested using the Mantel-Haenszel Chi-square test statistic:

$$\chi^2_{\mathrm{MHCC}} = \frac{(|\sum_i^K a_i - \sum_{i=1}^K E(a_i)| - 0.5)^2}{\sum_{i=1}^K V(a_i)}. \tag{3}$$

The expected number of correct male responses for a stratum is

$$E(a_i) = \frac{n_1 m_{1i}}{N_i}. \tag{4}$$

The Yates continuity correction factor of 0.5 in the numerator accounts for using a continuous χ^2 distribution to analyze a sample with a discrete distribution [47].

The variance within a stratum is

$$V(a_i) = \frac{n_1 n_2 m_{1i} m_{2i}}{N_i^2 (N_i - 1)}. \quad (5)$$

Various studies have shown conflicting results for the effects of sample size in the use of the Mantel-Haenszel procedure [48]; however, Fidalgo *et al.* [49] suggest that the procedure is appropriate for small sample sizes of 30 or more. In this work we have a large overall sample, $N = 1415$, with 1269 males in the reference group and 146 females in the focus group. We have used the criteria of Mantel and Fleiss [50] which uses the sum over all strata of the expectation values in each cell to ensure that there are enough students in each strata for the Mantel-Haenszel procedure to be valid [51].

Therefore, the level of bias in a question can be classified into three categories based on the absolute value of the log transformed Mantel-Haenszel odds ratio and whether that value is statistically significant using the level of significance or probability value p_{MH} [44,51,52]. In other words, the null hypothesis is defined by the odds ratio being zero at each stratum. We can flag a biased question if both conditions (i) $|\alpha^*_{\mathrm{MH}}| < 1$ and (ii) $p_{\mathrm{MH}} \leq 0.05$ are fulfilled.

Sometimes individual test statistics for multiple items are used to judge if the overall assessment is biased. In this situation, a statistical correction (e.g., Bonferroni) that applies a more stringent test of significance for each of the individual items may be needed [52,53]. In this study we are interested in possible bias in individual test questions and thus no correction is needed [54,55].

We can now place questions into the following categories.

- "none," where questions exhibit negligible difference between the odds ratios; α^*_{MH} is not significantly different from zero $|\alpha^*_{\mathrm{MH}}| < 1$ and $p_{\mathrm{MH}} > 0.05$
- "moderate," where questions exhibit a moderate but statistically significant difference; i.e., $1 < |\alpha^*_{\mathrm{MH}}| < 1.5$ and $p_{\mathrm{MH}} < 0.05$
- "strong," where questions exhibit a statistically significant and large difference; i.e., $|\alpha^*_{\mathrm{MH}}| \geq 1.5$ and $p_{\mathrm{MH}} < 0.05$.

III. RESULTS AND DISCUSSION

Figure 1 shows the α^*_{MH} parameter for the MH analysis divided into topics used in the multiple-choice examination. Approximately 80% of MCQs used in the final exam for Phys1210 show no statistically significant difference in gender performance over the eight-year period of this study as indicated by $|\alpha^*_{\mathrm{MH}}| < 1$, and thus are category A questions for the MH analysis. Overall, however, there is a small statistically significant gender bias to the complete set of multiple-choice exam questions. Analysis of the individual topics reveals that this bias varies from topic to topic. For example, two of the five wave physics and electricity

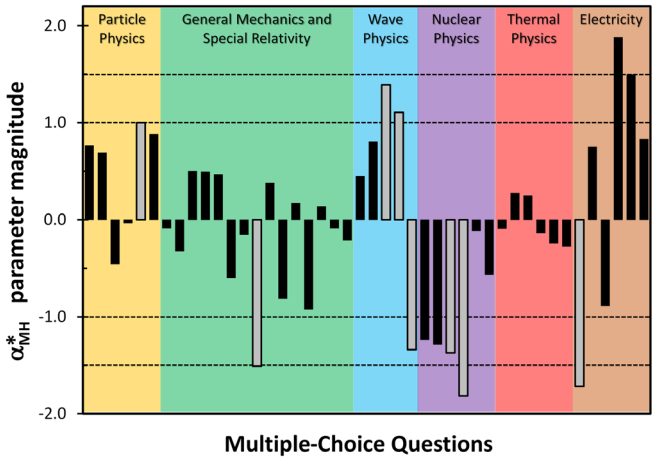

FIG. 1. Plot of the log transformed Mantel-Haenszel odds ratio (α^*_{MH}) parameter for each MCQ analyzed. The plot is split up into topics and shows gender relationships across these categories. A positive (negative) value of α^*_{MH} indicates a bias toward females (males), with a greater probability of a female (male) student answering the question correctly than a male (female) student of the same ability. The gray shaded bars show α^*_{MH} for which $p_{\mathrm{MH}} < 0.05$. The black shaded bars show α^*_{MH} for which $p_{\mathrm{MH}} > 0.05$.

questions are biased qualitatively toward female students ($\alpha^*_{\mathrm{MH}} > 1$; $p_{\mathrm{MH}} > 0.05$) while nuclear physics trends toward male biasing with 2 out of 6 questions exhibiting a statistically significant bias toward males ($\alpha^*_{\mathrm{MH}} < -1.5$; $p_{\mathrm{MH}} < 0.05$). In contrast to the other topics, the thermal physics section shows almost no variation in α^* parameter across all six questions ($|\alpha^*_{\mathrm{MH}}| < 0.27$). For first-year Australian University students, the thermal physics topic is typically the most difficult (since the content was generally not covered at high school level during this analysis period) with overall results that are lower than the other topics. We hypothesize that the lack of correct responses overall reduces the α^*_{MH} magnitude and thus no statistically significant trend to either bias is observed ($|\alpha^*_{\mathrm{MH}}| < 1$).

Overall, Fig. 1 shows that there were five questions that manifested with "moderate" biasing ($1 < |\alpha^*_{\mathrm{MH}}| < 1.5$ and $p_{\mathrm{MH}} < 0.05$) and three questions with "large" biasing ($|\alpha^*_{\mathrm{MH}}| \geq 1.5$ and $p_{\mathrm{MH}} < 0.05$).

Previous work has shown that the characteristics of assessment questions can result in gender bias [25]. Halpern *et al.* argued that males outperform females on visual-spatial questions whereas females tend to perform better on more "verbal" tasks [16]. More recently, a study of the impact of exam question structure on the performance of first-year physics undergraduates showed that while student performance improved with increased scaffolding of questions, the increase in average examination mark was greater for female students (13.4%) than for male students (9%) [34]. However, subsequent research

indicated that scaffolding was not the dominant determinant of gender gaps in MCQs, but instead that questions with a high visual-spatial content (diagrams and multidimensional context) were stronger indicators of male bias [35]. This result was consistent with work on the gender differences in performance over eight years in the Australian Science Olympiad Exam for physics, which revealed that the gender gaps in achievement correlated with the question type, particularly with respect to the content, context, and presentation [25].

In order to further probe the origin of the gender bias that is observed in our first-year MCQs, we developed a system for categorizing the questions designed to tease out the underlying characteristics that indicate gender bias in our questions. Our schema is based on several different categorization strategies of physics questions from the literature (see Refs. [26,35,56–59], and references therein). The assignment of categories was done separately of the questions being designated to the exam paper and was performed independently of the MH data analysis.

The following questions (listed as question one, two, three, and four) provide an illustration of the implementation of the categorization schema described in Table II.

TABLE II. Categorization schema used to characterize the MCQs.

Category	Question descriptor
Numbers (N)	Question or answers include numerical values whether they are needed for the MCQ solution or not.
Equation (E)	Requires an equation, from the equation list included in the exam, to answer the question. This category may be split into subcategories: single or multiple equation.
Words (W)	Comprises a large number of words in the answers. Based on analysis of the MCQ answers across all topics, the threshold was determined to be greater than 50 words.
Concept (C)	Involves a major physics concept that they must identify first before solving and not simply plug numbers into an equation. This category may be split into subcategories: single concept, multiple concept, and memory.
Image (I)	Includes an image of some description. Includes all diagrams, graphs, and schematics that must be interpreted to solve the question.
Visual (V)	Uses language that causes the reader to visualize or picture the problem or parts of the problem. Examples include push, rolling, collision, decay, conduction, heat, spaceship, placed. Note that both image and visualization categories can be used together.

Question one:

Which statement is false:

(a) *A body in uniform motion has no acceleration*
(b) *In circular motion with constant speed, the acceleration is perpendicular to the velocity*
(c) *Near the surface of the Earth, free fall is motion with constant acceleration*
(d) *When a body is in motion, a force must be acting on it*
(e) *When an object is in free fall, it is being accelerated*

Question one is from the general mechanics topic and involves (i) many words in the answers (58 words), (ii) a key physics concept (Newton's laws of motion), and (iii) language that requires visualization of the problem (a body in motion). Thus, this MCQ was categorized as containing the W, C, and V characteristics.

Question two:

Two point charges $X = +2$ μC *and* $Y = -3$ μC *are placed 100 mm apart. The electric potential V due to the two charges (along the line between them) will be zero at a distance, in mm, from X of:*

(a) *30*.
(b) *40*.
(c) *50*.
(d) *60*.
(e) *70*.

Question two is from the electricity topic and involves (i) numerical answers, (ii) an equation (electrostatic potential from a point charge), (iii) a key physics concept (conservative fields), and (iv) language that requires visualization of the problem (point objects separated by a known distance). Thus, this MCQ was categorized as containing the N, E, C, and V characteristics.

Question three:

If an object satisfying Hubble's Law is 10^9 *parsecs away, how fast is it travelling?*

(a) *0.99 c*
(b) *0.8 c*
(c) *0.5 c*
(d) *0.43 c*
(e) *0.23 c*

Question three is from the particle physics topic and involves (i) numerical answers and (ii) a single equation (Hubble's law). Thus, this MCQ was categorized as containing the N and E characteristics.

Question four:

A stretched cord, fixed at both ends, vibrates at a frequency of 12 Hz with a standing transverse wave pattern as shown. What would be the frequency of the third harmonic?

(a) *9 Hz*
(b) *12 Hz*
(c) *18 Hz*
(d) *24 Hz*
(e) *36 Hz*

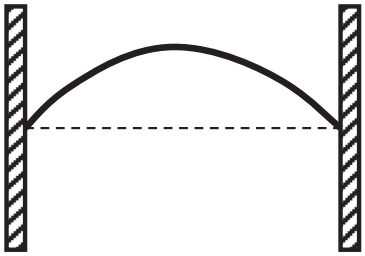

FIG. 2. The image attached to question four.

Question four is from the wave physics topic and includes the diagram in Fig. 2. This questions involves: (i) numerical answers, (ii) an equation (wave equation), (iii) a key physics concept (standing waves), (iv) an image, and (v) language that requires visualization of the problem (vibrating stretched cord). Thus, this MCQ was categorized as containing the *N*, *E*, *C*, *I*, and *V* characteristics.

Figure 3 shows the frequency distribution of categories for all the questions analyzed within this study. As can be seen in the figure, the visual, concept, equation, and numbers categories dominate the characteristics, while only a few questions contain either images or have answers that comprise a large number of words.

As a further illustration, the MH analysis of the example questions one to four is shown in Table III, revealing that all show either male or female bias. Indeed, for the eight questions that show gender bias, five are male biased and three are female biased. Interestingly, all the category *C* questions exhibit male bias.

Figure 4 shows the frequency distribution of categories for the MCQs that exhibit male or female bias, plotted as a percentage of the total number of male or female biased questions, respectively. For the analysis presented here, the single or multiple equation and memory, single, or multiple concept subcategories were consolidated into a single equation and concept category, respectively.

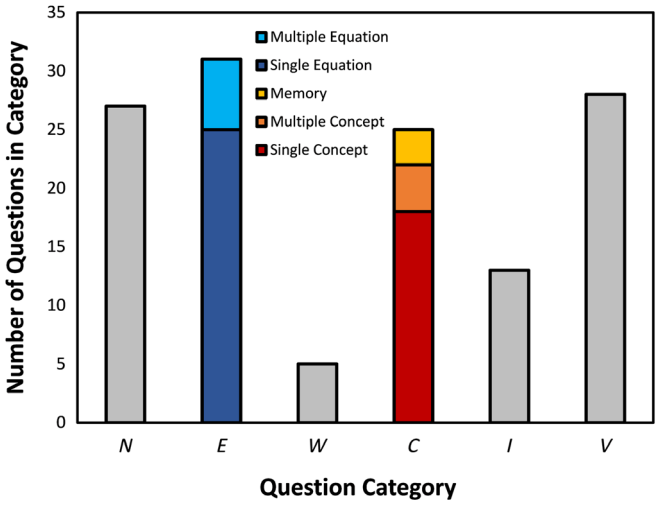

FIG. 3. The number of times each characteristic is observed within the set of questions used in the first-year advanced physics exam.

TABLE III. MH analysis characteristics of the example MCQs one to four.

Question	α^*_{MH}	p_{MH}	Bias	Bias category
One	−1.51	0.01	Male	Strong
Two	−1.71	0.05	Male	Strong
Three	1.00	0.04	Female	Moderate
Four	1.11	0.02	Female	Moderate

Although the sample size is small, there are some trends in the relative distribution of the characteristics of gender biased questions. Figure 4 reveals that MCQs exhibiting male bias are dominated by the *C* and *V* characteristics. Comparing the ratio of the probability that any question exhibiting gender bias (i.e., combining category "moderate" and "strong" questions) contains the *V* and *C* characteristics (0.63) with the probability that the *V* and *C* characteristics are present in a nongender biased question (0.28) revealed that it was 2.3 times more likely that these questions contain visualization and concept characteristics than questions that show no statistically significant difference in gender. If only male biased questions are considered, the probability ratio rises to 2.8, whereas for female questions the probability ratio is only 1.1, indicating that there is no real tendency for female biased questions to contain the *V* and *C* characteristics.

As discussed earlier, visualization as a cause of gender bias toward males in physics exam questions has been identified in several publications (see Refs. [7,15,56,57], and references therein). Several studies have investigated gender bias in the Force Concept Inventory which includes

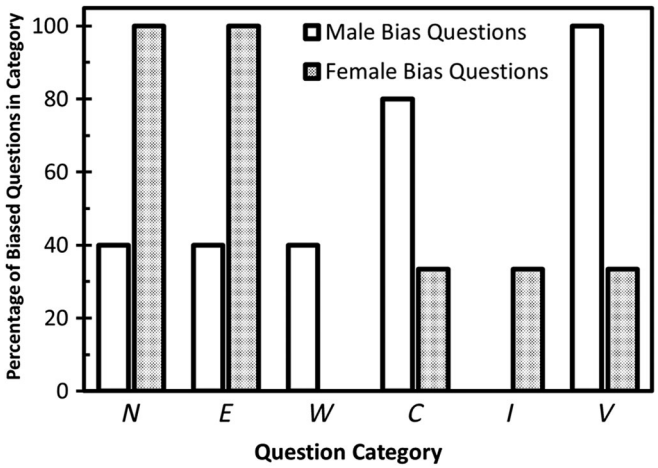

FIG. 4. Comparison of the percentage distribution of characteristics for male and female biased questions. Out of the eight questions that show gender bias, five are male biased and three are female biased. For example, for the numbers category (*N*), the bars show that 40% of the male bias questions are category *N* and 100% of the female bias questions are category *N*.

many visual type questions, such as projectile motion and free-body diagrams [7,15]. These studies showed that male biased questions included a diagram and involved spatial and grammar cues to visualize motion. This observation correlates well with the studies of male brain activity revealing increased aptitude toward spatial and abstract mechanical concepts. [59] Both Dawkins *et al.* [35] and Wilson *et al.* [26] analyzed a variety of exam questions and revealed large gender gaps for questions that include an image or diagram as well as a multidimensional context. In contrast to the Dawkins and Wilson studies, our analysis characterized images separately from grammar induced visualization and found no male bias, suggesting that the inclusion of an image may reduce the tendency for male bias in visualization-based questions. Indeed, Chen *et al.* [58] showed that the inclusion of diagrams in physics-based questions assisted students in choosing the correct concept to answer the question. Further work is required to test this hypothesis more fully.

Returning to Fig. 4, we see that MCQs exhibiting female bias appear more likely to have the *N* and *E* characteristic. Moreover, two of the three questions which show moderate bias toward female students are from the wave physics section. However, unlike the situation with the male biased questions, this behavior does not appear to correlate with previous studies. For example, Wilson *et al.* found a large female bias in two similar MCQs involving conservation of momentum [26]. Most studies have found that on average female students tend to outperform male students on verbal and reading tests. However, the *W* characteristic does not appear to be an indicator for female bias. Previous work has indicated that female students could sustain their performance on tests and that longer tests reduce male bias [60–62]. Given that our exam is three hours long, the fact that the wave physics sections occur toward the end of the examination paper may contribute to an increased tendency toward female bias in these questions.

IV. LIMITATIONS AND FUTURE DIRECTIONS

This work represents an initial attempt toward the development of such a novel gender bias identification system for MCQ exams for university-level physics teachers. The fact that only a few biased questions have been identified means that further work is required before generalizations can be made. However, the identification of biased questions based on simple categories is an important first step. Future work will seek to increase the sample size by exploring gender bias in exam performance for first-year physics subjects across multiple institutions. It is known that changing the way information is presented and adjusting question context can eliminate performance bias in MCQ tests [25]. As such, future study will explore how the *NEWCIV* characterization system presented in this paper could offer the prospect of a method for prior identification and mitigation of gender bias rather than postevaluation of its consequences. As part of future work, a selection of students post exam will be asked to characterize the multiple-choice questions to enquire whether there is a difference between student and academic process. Students will also be asked to verbally interpret a set of questions and describe their chosen answer and their reasoning for it to possibly find out whether there are gender differences in student approaches to given questions.

There are a variety of university-based guidelines to assist academics in choosing the optimal assessment for their material as well as providing support for creating a selection of question types that evaluate different learning outcomes [63]. For multiple-choice questions this guidance includes, for example, avoiding long complex sentences and ambiguous language [64–66]. Although some of these guidelines include a selection of evaluation tools that value diversity of learning needs, there is little specific information on how to identify and address bias in examination questions [67]. While MCQs are not inherently biased, studies have identified male performance bias in physics-based MCQ exams and yet there is currently no widely available benchmark tool to test for gender bias in these assessments [24,35,36].

V. CONCLUSION

The presence of gender bias in first-year multiple-choice physics examinations has been investigated. In summary, the MH test is applicable for minimum sample sizes as small as 30 and is the test of choice as it effectively eliminates student ability as a variable in determining gender bias in multiple-choice examinations. The study involved assessing the individual performance of over 1400 students over an eight-year period and revealed that approximately 20% of the questions exhibit some form of statistically significant bias, with 12.5% of questions biased toward males and 7.5% of questions biased toward females. In order to further understand the origin of the observed bias, each question was categorized in terms of a schema designed to determine the key characteristics of each question. The analysis revealed that the questions exhibiting male bias were more likely to contain both the visualization and concept characteristics, consistent with previous studies showing visualization as a cause of gender bias toward males in physics exam questions. By contrast, questions exhibiting female bias were not correlated with the words characteristic (indicated by previous work as a potential marker for female bias) but instead were more likely to contain both the numbers and equation characteristics. This work highlights that analysis of gender bias in multiple-choice physics examinations (based on question characteristics) might be a useful tool in understanding the presence and origin of gender gaps in student performance.

ACKNOWLEDGMENTS

The authors would like to thank the University of Newcastle for financial assistance. This work comes under the Human Research Ethics committee (HREC) Quality Assurance proposal QA92.

[1] E. A. Dare and G. H. Roehrig, "If I had to do it, then I would": Understanding early middle school students' perceptions of physics and physics-related careers by gender, Phys. Rev. Phys. Educ. Res. **12**, 020117 (2016).

[2] J. Day, J. B. Stang, N. G. Holmes, D. Kumar, and D. A. Bonn, Gender gaps and gendered action in a first-year physics laboratory, Phys. Rev. Phys. Educ. Res. **12**, 020104 (2016).

[3] R. Ivie, S. White, and R. Y. Chu, Women's and men's career choices in astronomy and astrophysics, Phys. Rev. Phys. Educ. Res. **12**, 020109 (2016).

[4] F. Siddiq and R. Scherer, Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy, Educ. Res. Rev. **27**, 205 (2019).

[5] Office of the Chief Scientist, Science, Technology, Engineering and Mathematics: Women in STEM, Office of the Chief Scientist, Canberra, 2016, viewed 18 January 2019, https://www.chiefscientist.gov.au/sites/default/files/OCS_Women_in_STEM_datasheet.pdf.

[6] A. Binnie, Development of a senior physics syllabus in New South Wales, Phys. Educ. **39**, 490 (2004).

[7] J. Docktor and K. Heller, Gender differences in both Force Concept Inventory and introductory physics performance, AIP Conf. Proc. **1064**, 15 (2008).

[8] S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch, and N. R. Walet, Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison, Eur. J. Phys. **34**, 421 (2013).

[9] S. J. Pollock, Comparing student learning with multiple research-based conceptual surveys: CSEM, and BEMA, AIP Conf. Proc. **1064**, 171 (2008).

[10] Department of Education and Training, uCube—Higher Education Data Cube, Department of Education and Training, Canberra, viewed 7 November 2021, https://highereducationstatistics.education.gov.au/.

[11] National Centre for Vocational Education Research, VOCSTATS Resources, extracted on 4 October 2018, https://www.industry.gov.au/publications/advancing-women-stem-strategy/snapshot-disparity-stem.

[12] Advancing Women in STEM strategy: Snapshot of disparity in STEM, https://www.industry.gov.au/data-and-publications/advancing-women-in-stem-strategy/snapshot-of-disparity-in-stem, retrieved 31 March 2022.

[13] H. Pinson, Y. Feniger, and Y. Barak, Explaining a reverse gender gap in advanced physics and computer science course-taking: An exploratory case study comparing Hebrew-speaking and Arabic-speaking high schools in Israel, J. Res. Sci. Teach. **57**, 1177 (2020).

[14] S. Moshfeghyeganeh and Z. Hazari, Effect of culture on women physicists' career choice: A comparison of Muslim majority countries and the West, Phys. Rev. Phys. Educ. Res. **17**, 010114 (2021).

[15] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory?, AIP Conf. Proc. **1413**, 171 (2012).

[16] D. F. Halpern, C. P. Benbow, D. C. Geary, R. C. Gur, J. S. Hyde, and M. A. Gernsbacher, The science of sex differences in science and mathematics, Psychol. Sci. Publ. Interest **8**, 1 (2007).

[17] S. J. Ceci, W. M. Williams, and S. M. Barnett, Women's underrepresentation in science: Sociocultural and biological considerations, Psychol. Bull. **135**, 218 (2009).

[18] S. Thompson, L. De Bortoli, C. Underwood, and M. Schmid, PISA in brief I: Student performance, Australian Council for Educational Research (2019), https://research.acer.edu.au/ozpisa/34/.

[19] J. Cromley, Reading achievement and science proficiency: International comparisons from the Programme on International Student Assessment, Read. Psychol. **30**, 89 (2009).

[20] T. O'Reilly and D. S. McNamara, The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement, Am. Educ. Res. J. **44**, 161 (2007).

[21] N. Cruz Neri, K. Guill, and J. Retelsdorf, Language in science performance: Do good readers perform better?, Eur. J. Psychol. Educ. **36**, 45 (2021).

[22] R. B. Prophet and N. B. Badede, Language and student performance in junior secondary science examinations: The case of second language learners in Botswana, Int. J. Sci. Educ. **7**, 235 (2009).

[23] C. Rivera and C. W. Stansfield, The effect of linguistic simplification of science test items on score comparability, Educ. Assess. **9**, 79 (2004).

[24] E. Bird and G. Welford, The effect of language on the performance of second-language students in science examinations, Int. J. Sci. Educ. **17**, 389 (1995).

[25] J. R. T. Cassels and A. H. Johnstone, The effect of language on student performance on multiple choice tests in chemistry, J Chem. Educ. **61**, 613 (1984).

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

[27] R. Henderson, G. Stewart, J. Stewart, L. Michaluk, and A. Traxler, Exploring the gender gap in the conceptual survey of electricity and magnetism, Phys. Rev. Phys. Educ. Res. **13**, 020114 (2017).

[28] K. N. Quinn, M. M. Kelley, K. L. McGill, E. M. Smith, Z. Whipps, and N. G. Holmes, Group roles in unstructured

labs show inequitable gender divide, Phys. Rev. Phys. Educ. Res. **16**, 010129 (2020).
[29] G. C. Marchand and G. Taasoobshirazi, Stereotype threat and women's performance in physics, Int. J. Sci. Educ. **35**, 3050 (2013).
[30] C. R. Fisher, R. H. Brookes, and C. D. Thompson, 'I don't study physics anymore': A cross-institutional Australian study on factors impacting the persistence of undergraduate science students, Res. Sci. Educ. **52**, 1565 (2022).
[31] M. Løken, When research challenges gender stereotypes: Exploring narratives of girls' educational choices, in *Understanding Student Participation and Choice in Science and Technology Education* (Springer, Dordrecht, 2015), pp. 277–295.
[32] Z. Y. Kalender, E. Marshman, C. Schunn, T. Noeks-Malach, and C. Singh, Framework for unpacking students' mindsets in physics by gender, Phys. Rev. Phys. Educ. Res. **18**, 010116 (2022).
[33] A. L. Traxler, X. C. Cid, J. Blue, and R. Barthelemy, Enriching gender in physics education research: A binary past and complex future, Phys. Rev. Phys. Educ. Res. **12**, 020114 (2016).
[34] V. Gibson, L. Jardine-Wright, and E. Bateman, An investigation into the impact of question structure on the performance of first year physics undergraduate students at the University of Cambridge, Eur. J. Phys. **36**, 045014 (2015).
[35] H. Dawkins, H. Hedgeland, and S. Jordan, Impact of scaffolding and question structure on the gender gap, Phys. Rev. Phys. Educ. Res. **13**, 020117 (2017).
[36] H. Hedgeland, H. Dawkins, and S. Jordan, Investigating male bias in multiple choice questions: Contrasting formative and summative settings, Eur. J. Phys. **39**, 055704 (2018).
[37] S. F. Reardon, D. Kalogrides, E. M. Fahle, A. Podolsky, and R. C. Zárate, The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eighth grades, Educ. Res. **47**, 284 (2018).
[38] S. Salehi, S. Cotner, S. M. Azarin, E. E. Carlso, M. Driessen, V. E. Ferry, W. Harcombe, S. McGaugh, D. Wassenberg, A. Yonas, and C. J. Ballen, Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety, Front. Educ. **4**, 107 (2019).
[39] M. Dew, J. Perry, L. Ford, W. Bassichis, and T. Erukhimova, Gendered performance differences in introductory physics: A study from a large land-grant university, Phys. Rev. Phys. Educ. Res. **17**, 010106 (2021).
[40] C. Castellanos-Serrano, G. Escribano, J. Paredes-Gázquez, and E. San-Martin Ganzález, What is behind the gender gaps in economics distance education: Age, work-life balance and Covid-19, PLoS One **17**, e0272341 (2022).
[41] E. M. Aucejo, J. French, M. P. Ugalde Araya, and B. Zafar, The impact of Covid-19 on the student experiences and expectations: Evidence from a survey, Journal of public economics **191**, 104271 (2020).
[42] G. Casalone, A. Michelangeli, J. Osth, and U. Turk, The effect of lockdown on students' performance: A comparative study between Sweden, Italy and Turkey, Heliyon **9**, e16464 (2021).
[43] G. Orlov, D. McKee, J. Berry, A. Boyle, T. DiCiccio, T. Ransom, Z. R. Patterson, and R. J. McQuaid, Coping with the Covid-19 pandemic: Examining gender differences in stress and mental health among university students, Front. Psychol. **12**, 439 (2021).
[44] P. Phanphech, T. Tanitteerapan, N. Mungkung, S. Arunrungrusmi, C. Chunkul, A. Songruk, T. Yuji, and H. Kinoshita, An analysis of student anxiety affecting online learning on conceptual applications in physics: Synchronous vs. asynchronous learning, Educ. Sci. **12**, 278 (2022).
[45] N. Mantel and W. Haenszel, Statistical aspects of the analysis of data from retrospective studies of disease, J. Natl. Cancer Inst. **22**, 719 (1959).
[46] P. Martinková, A. Drabinová, Y. L. Liaw, E. A. Sanders, J. L. McFarland, and R. M. Price, Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments, CBE Life Sci. Educ. **16**, rm2 (2017).
[47] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions* (John Wiley & Sons, Inc., New York, 2003).
[48] D. Emily, G. Brooks, and G. Johanson, Detecting differential item functioning: Item response theory methods versus the Mantel-Haenszel procedure, Int. J. Assess. Tools Educ. **8**, 376 (2021).
[49] Á. M. Fidalgo, K. Hashimoto, D. Bartram, and J. Muñiz, Empirical Bayes versus standard Mantel-Haenszel statistics for detecting differential item functioning under small sample conditions, J. Exp. Educ. **75**, 293 (2007).
[50] N. Mantel and J. L. Fleiss, Minimum expected cell size requirements for the Mantel-Haenszel one-degree-of-freedom chi-square test and a related rapid procedure, Am. J. Epidemiol. **112**, 129 (1980).
[51] A. M. Fidalgo, D. Ferreres, and J. Muniz, Utility of the Mantel-Haenszel procedure for detecting differential item functioning in small samples, Educ. Psychol. Meas. **64**, 925 (2004).
[52] M. Zieky, *A DIF Primer* (Educational Testing Service, Princeton, NJ, 2003).
[53] S. Lee and D. K. Lee, What is the proper way to apply the multiple comparison test?, Korean J. Anesthesiol. **71**, 353 (2018).
[54] R. A. Armstrong, When to use the Bonferroni correction, Ophthalmic Physiolog. Opt. **34**, 502 (2014).
[55] T. V. Perneger, What's wrong with Bonferroni adjustments, Br. Med. J. **316**, 1236 (1998).
[56] D. J. Low and K. F. Wilson, Persistent gender gaps in first-year physics assessment questions, in *Proceedings of the Australian Conference on Science and Mathematics Education, Curtin University*, 2015, ISBN 978-0-9871834-4-6, pp. 118–124.
[57] L. J. Rennie and L. H. Parker, Equitable measurement of achievement in physics: High school students' responses to assessment tasks in different formats, and contexts, J. Women Minorities Sci. Eng. **4**, 113 (1998).
[58] Q. Chen, G. Zhu, Q. Liu, J. Han, Z. Fu, and L. Bao, Development of a multiple-choice problem-solving categorization test for assessment of student knowledge

structure, Phys. Rev. Phys. Educ. Res. **16**, 020120 (2020).
[59] S. Siddiqui, Categorised and correlated multiple-choice questions: A tool for assessing comprehensive physics knowledge of students, Educ. Sci. **12**, 575 (2022).
[60] M. S. Gurian and K. Stevens, With boys and girls in mind, Educ. Leader. **62**, 21 (2004).
[61] S. M. Lindberg, J. S. Hyde, J. L. Petersen, and M. C. Linn, New trends in gender and mathematics performance: A meta-analysis, Psychol. Bull. **136**, 1123 (2010).
[62] T. S. Dee, Teachers and the gender gaps in student achievement, J. Hum. Resour. **42**, 528 (2007).
[63] P. Balart and M. Ooesterveen, Females show more sustained performance during test-taking than males, Nat. Commun. **10**, 3798 (2019).
[64] https://ldti.newcastle.edu.au/post/assessment-task-activities.
[65] J. R. Godfrey, Cunningham, George K. (1998). Assessment in the classroom: Constructing and interpreting tests. London: Falmer Press. vii + 225 pages, Aust. J. Teach. Educ. **23**, 5 (1998).
[66] A. W. Ward and M. Murray-Ward, *Assessment in the Classroom* (Wadsworth Publishing Co, Belmont, CA, 1999).
[67] https://policies.newcastle.edu.au/document/view-current.php?id=137.