Differences in gender performance on competitive physics selection tests

Kate Wilson*
School of Engineering and Information Technology, UNSW Canberra at the Australian Defence Force Academy, Canberra BC, ACT 2610, Australia

David Low
School of Physical, Environmental and Mathematical Sciences, UNSW Canberra at the Australian Defence Force Academy, Canberra BC, ACT 2610, Australia

Matthew Verdon†
Australian Science Olympiads Physics Program, Australian Science Innovations, Canberra, Australia and ASMS, Flinders University, Sturt Rd, Bedford Park, Adelaide 5042, Australia

Alix Verdon
Australian Science Olympiads Physics Program, Australian Science Innovations, Canberra, Australia
(Received 29 January 2015; published 1 August 2016)

[This paper is part of the Focused Collection on Gender in Physics.] We have investigated gender differences in performance over the past eight years on the Australian Science Olympiad Exam (ASOE) for physics, which is taken by nearly 1000 high school students each year. The ASOE, run by Australian Science Innovations (ASI), is the initial stage of the process of selection of teams to represent Australia at the Asian and International Physics Olympiads. Students taking the exam are generally in their penultimate year of school and selected by teachers as being high performing in physics. Together with the overall differences in facility, we have investigated how the content and presentation of multiple-choice questions (MCQs) affects the particular answers selected by male and female students. Differences in the patterns of responses by male and female students indicate that males and females might be modeling situations in different ways. Some strong patterns were found in the gender gaps when the questions were categorized in five broad dimensions: content, process required, difficulty, presentation, and context. Almost all questions saw male students performing better, although gender differences were relatively small for questions with a more abstract context. Male students performed significantly better on most questions with a concrete context, although notable exceptions were found, including two such questions where female students performed better. Other categories that showed consistently large gaps favoring male students include questions with projectile motion and other two-dimensional motion or forces content, and processes involving interpreting diagrams. Our results have important implications, suggesting that we should be able to reduce the gender gaps in performance on MCQ tests by changing the way information is presented and setting questions in contexts that are less likely to favor males over females. This is important as MCQ tests are frequently used as diagnostic tests and aptitude tests as well as to assess learning.

DOI: 10.1103/PhysRevPhysEducRes.12.020111

I. MOTIVATION

Each September, nearly 1000 Australian high school students take the Australian Science Olympiad Exam (ASOE) for physics. On the basis of performance in the ASOE, approximately 24 students are chosen to attend a summer school "training camp" in January of the following year. The top eight students from the summer school make up the Australian team for the Asian Physics Olympiad (APhO) in April of that year, and five of these students go on to attend the International Physics Olympiad (IPhO) in July. The ASOE is the thus the first hurdle in an intensive selection and training process.

The students who take the ASOE are selected by their teachers as being of high ability in physics, and are mainly in year 11 (the penultimate year of high school) and about 16 years old. Typically between a quarter and a third of these students are female, which is approximately the same fraction as the total year 11 physics cohort. However, the fraction of females who are offered places at the summer school is consistently lower than this, seldom more than five out of the 24.

The gender difference in ASOE performance was observed as an ongoing issue in the selection process as early as 2005 [1], and modifications were made to the

*k.wilson@adfa.edu.au
†matthew.verdon@asi.edu.au

ASOE in an effort to address this perceived inequity. Before 2007, the ASOE involved 20 multiple-choice questions (MCQs) and four written-answer questions which were typically of a mathematical nature. In 2007, the number of MCQs on the test was halved, as it has often been observed that female students do not perform as well as male students on MCQ diagnostic tests (see, e.g., Refs. [2,3]). The variety of content addressed in the MCQs was increased, and a wider range of contexts was employed. In the written section, candidates were allowed to choose four questions to answer from a selection of five or six, and an extended range of skills was addressed. The option of choice was removed a few years later when it was found to be ineffective, but the written questions address a further extended range of skills.

The MCQs typically include conceptual questions similar to those on the Force Concept Inventory (FCI) [4,5], scientific reasoning questions, and "process questions" (such as the calculation and expression of uncertainties). The written section includes questions requiring algebraic manipulations and calculations, the graphing and interpretation of data, written explanations of physical phenomena, and experimental design. The questions explore a mix of topics and content that should be both familiar and new to the students. (See Ref. [6] for past ASOE papers.) A new paper is set every year, although some questions are recycled from time to time.

Despite these changes, which were informed by the literature at the time dealing with gender differences in physics assessment, no significant, repeatable decrease in gender gap in performance on the exam, or any increase in number of female students selected to continue beyond this stage, was achieved. The analysis of ASOE data from 2007 to 2014 shows that issues surrounding gender performance remain. For almost all questions, the gender gap is such that males outperform females. In addition, there are gender differences in the patterns of answer selection in the MCQ questions.

We see similar differences in the performance of female and male students in the written section of the ASOE. This section is closer in style to the questions which students would encounter in the APhO and IPhO.

In terms of the marks given, the written section of the ASOE is weighted at about 4 times that of the MCQ section. However, students' marks in the written section are, in absolute terms, on average only about 1.5 times their mark in the MCQ section (i.e., students find the written section significantly more difficult), and the gender gap in results is not reduced. Hence, the issues observed and discussed here are not simply a matter of a gender difference between MCQs and written answers: female students are not performing as well as male students on both sections of the ASOE.

Given these observations, in this paper we will explore only the MCQ results. This removes the need for us to classify student written answers according to misconception or approach, particularly when the marks in the written section are so low and therefore less amenable to detailed analysis. This paper, then, sets out to detail and explore the gender differences in the MCQ section of the ASOE, and attempts to identify the origin of some of these differences.

II. BACKGROUND

Studies of gender differences as related to ability and achievement are widespread in the literature. Particular attention has been paid to the topic since the seminal (and polarizing) work of Maccoby and Jacklin [7], and it continues to draw high-level attention as countries improve their understanding of the implications of disenfranchising a significant segment of their population (see, e.g., Refs. [8,9]). The main questions surrounding observed gender differences—whether they are innate or acquired, whether they apply regardless of context, and whether they can be overcome by teaching and/or experience—remain open, and the answers often vary depending on the study and the particular field of inquiry.

The comprehensive review by Halpern *et al.* [10] covers the preceding three decades of work on gender differences in the sciences and mathematics. In summary, and noting that there are many complex links between contributing factors, and many caveats when it comes to generalizations, females tend to excel in "verbal" activities, while males exhibit better performance in "visual-spatial" tasks. However, males also tend to exhibit more variability within the cohort: there are relatively more males at both ends of the achievement distribution, and the proportion of males at the high-achievement end tends to increase with age. Starting from the high school years, males also appear to perform better in tasks that require knowledge to be applied in a problem-solving or "real-world" context [3,11].

There do appear to be physiological gender differences, as revealed by scans of brain activity in adolescents [12]. Girls tend to use more cortical areas for verbal functions, while boys use this area more for abstract and physical-spatial functions. This trend makes boys more comfortable moving things through space, and better suited to using diagrams and pictures, while girls are better at multitasking, concentrating, and reading. McBride [13] takes this further, observing that girls are better at discriminating objects (e.g., "what is it?"), while boys are better at location and movement (e.g., "where is it?"). He notes that girls deal better with the complexities of reading than boys, who have a tendency to lose interest in a problem if the instructions are layered too deeply.

There have been many studies of gender-based differences in physics, with most concentrating on the relative performance of males and females in standardised tests such as the FCI. A detailed review of this literature is presented by Madsen *et al.* [14], who conclude that

observed gaps are likely to be a combination of many small factors, and suggest that isolated explanations need to be treated with caution due to a lack of repeatability. One common theme, however, appears to be females underperforming relative to males on questions that involve vertical and/or two-dimensional motion (see, e.g., Refs. [15–17]). If "facility" is defined as "the fraction of a cohort that answers a question correctly," typical gender gaps (defined as the difference between the male and female facilities) on FCI questions range between −0.10 and +0.36, where a positive number indicates a performance bias in favor of males.

Two main criticisms have been directed at the FCI in terms of how it might be biased against female students. First, in terms of the wording and setting of some questions, the idea that females prefer problems with a "concrete" or "real-world" rather than "abstract" context [18–20] led McCullough [21] to rewrite the FCI using (stereotypical) female contexts with concrete settings. For example, a cannonball fired from a cliff became a bowl falling off a table, rolling steel balls became oranges, a ball in a circular channel became a girl on a water slide, etc. While the results indicated that context did interact with gender and affected performance so as to reduce the gender gap, it was generally a case of females performing no better but males performing worse in the rewritten version. Similarly mixed results from a context-based investigation of the FCI have been seen by others [2,22–24].

Second, the FCI is composed of 30 MCQs, and MCQs themselves have been identified by a number of studies as being problematic for females. This may be due to a tendency for males to take a black-and-white view and decide on the correct answer, perhaps via a strategic or elimination-based approach, while females may be more likely to see ambiguity in the proffered options and consider how each might be correct by looking for commonality [2,3,25–30].

Thus, there are a number of possible causes for observed differences in performance by males and females on MCQ-based physics questions. In what follows, we place the Australian Physics Olympiad ASOE results into the context described by the existing literature, and provide interpretation of the gender differences that we have observed.

III. DATA SET

The data set that we are examining is the results of the 2007–2014 physics ASOEs. Over these eight years, approximately 7000 high school students took these exams. Typically, only a few students from a given class are nominated by their teacher to attempt the ASOE, as it is aimed at only the most able physics students. The ASOE is aimed at students in year 11, as they will be in the final year of high school in the following year (students may not compete in the International Physics Olympiad competition once they have begun a university degree). Year 11 is also the first year where physics is offered as a stand-alone course in Australian high schools: before that, it is a component of a general science course. Teachers are asked to nominate only year 11 students, and in exceptional circumstances, year 10 students; however, there are invariably some students from other years nominated. Given these reasons, and noting that over 80% of students attempting the ASOE are in year 11, for consistency we restrict our analysis in this paper to the year 11 cohort.

The ASOE participation details are presented in Table I. In each of the years for which data are presented, the total number of students taking year 11 physics across Australia was approximately 30 000 [31], so approximately 2% of the entire cohort attempt the ASOE. Females typically made up approximately 26% of the year 11 and 12 physics cohort between 2007 and 2014. Amongst the students undertaking the ASOE, there were between 26% and 32% females between 2007 and 2014 (gender is self-identified by candidates attempting the ASOE). In the absence of supporting data, we could speculate that this difference might be due to teachers choosing some female students to participate in the ASOE at least partly on equity grounds rather than purely on ability. If this is the case, any bias in

TABLE I. The numbers of (self-identified) male and female students in each cohort. The total is slightly less than the number of students who took the exam, as not all students identified their gender. Figures in parentheses include students in years other than year 11.

Year	2007	2008	2009	2010	2011	2012	2013	2014	Total
No. of male									
Year 11	591	539	562	439	516	404	483	475	4003
(Total)	(695)	(662)	(702)	(574)	(619)	(528)	(618)	(634)	(5023)
No. of female									
Year 11	212	241	248	172	204	189	174	210	1647
(Total)	(242)	(292)	(285)	(211)	(252)	(234)	(224)	(283)	(2020)
% of female									
Year 11	26.4	30.9	30.6	28.2	28.3	31.9	26.5	30.7	29.2
(Total)	(25.8)	(30.6)	(28.9)	(26.9)	(28.9)	(30.7)	(26.6)	(30.9)	(28.7)

our analysis is likely to be small, given the numbers involved. It is also quite possible that the criteria on which teachers are judging "physics ability" are quite different from those tested for in the ASOE, for example, a whole-course performance approach.

A. General characteristics of the data set

The facility (fraction of students answering correctly) was calculated for each question, for the entire cohort as well as for male and female subsets. The difference in facility ([male facility] − [female facility]) was calculated for each question. The distribution of answers selected by males and females for every question was recorded. Gender-specific item response curves [32] were plotted for each question from the distribution of answer choices.

The total question set consists of 80 MCQs, although some questions are repeated with slight changes of context. The most repeated question is question 15 (Q15) from the 1995 version of the FCI [5], in which a car pushes a truck and students need to use Newton's third law to compare the force that each vehicle exerts on the other. This question was used five times in the eight years analyzed here, with the context as a car pushing a truck, a bus, or a train, and a tugboat pushing a ship. Other repeated questions are typically used only twice in this period.

The average mark out of 10 on the MCQ section for the eight years examined was 5.28 (s.d. 2.06) for males and 4.68 (s.d. 2.01) for females. This difference is significant at the 1% level (Cohen's $d = 0.29$). Figure 1 shows a histogram of male and female marks for the entire data set. Note that the distribution of marks for females is skewed towards lower marks, compared to the males. There are only a very small number of females achieving high marks. Females make up a total of 29%, or a little under one in three of the cohort. However, they make up only one in seven of the students achieving full marks, and one in six of those achieving a score of nine out of ten. Given that only those students who achieve in the top 2.5% overall are invited to continue with the team selection process, this gap in performance at the higher end of the MCQ marks contributes significantly to the low rate of selection of females for the summer school.

Gender gaps ([average male facility] - [average female facility]) on individual questions vary from 0.10 in favor of females to 0.28 in favor of males, with males outperforming females on the majority of questions. The average gap over all 80 questions was 0.06 in favor of males. Over the ten multiple-choice questions, this means an average mark difference of 0.6/10, as seen in the average marks given above. The distribution of gaps is shown in Fig. 2.

The significance of the gap on any particular question depends on the facility of the question and the sample size. In what follows, we have checked for the significance of any difference in facility by gender via a chi-square test for homogeneity (with the two genders one way, and correct or incorrect the other way, giving one degree of freedom), and quote the χ^2 value where appropriate. Under this approach, a χ^2 value greater than 3.841 (6.635) is significant at the 5% (1%) level. As a general guide, however, a gap larger than 0.07 is significant at the 5% level for most questions. Throughout this paper, we use "large" and "small" to refer to the absolute magnitude of gaps, relative to the average gap in facility. Large gaps are typically greater than 0.10, while small gaps are typically less than 0.05. Standardized effect sizes in the comparison of gaps on individual questions are parametrized by the ϕ coefficient ($\phi = \sqrt{\chi^2/N}$). While the absolute magnitude of effect sizes in this paper are not numerically large, one is reminded that effect sizes depend on the area of investigation, the context, and the research method [33]. Since gender variation is secondary to the individual variation within the population, effect sizes cannot be numerically large [34]. Hence, following the recommendations of Baguley [35], in this paper we quote both the magnitude of the gender gap (the "simple" effect size) along with the "standardized" effect size ϕ, and encourage other researchers in the field to do the same, as an aid to comparison.

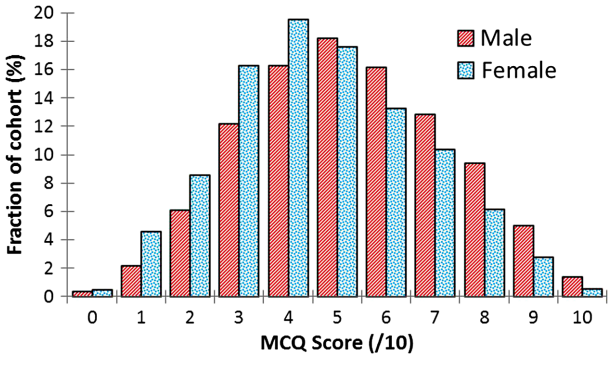


FIG. 1. Frequency distribution of male (red or striped) and female (blue or dotted) MCQ score (out of ten) for all ASOEs from 2007 to 2014.

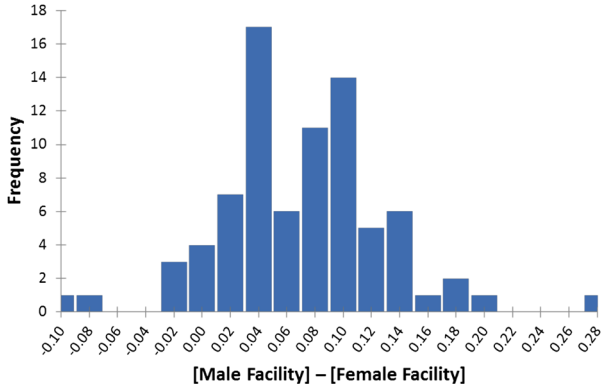


FIG. 2. Distribution of gender gaps ([male facility]-[female facility]) on all ASOE MCQs from 2007 to 2014.

IV. CATEGORIZATION OF QUESTIONS

While the data presented in Figs. 1 and 2 give a clear picture of a gap in performance between males and females, they do not give us any insight into the causes of the gaps. To investigate this, the ASOE MCQ questions were sorted into categories in five broad dimensions: content, process required, difficulty, presentation, and context. These five dimensions were chosen after initial simpler classification schemes with fewer dimensions were discarded when no clear patterns in gender gaps emerged using those schemes.

These five dimensions were arrived at based on previous categorizations described in the literature (for example, Hazel *et al.* [2], Halpern *et al.* [10], Gurian and Stevens [12], McBride [13], McCullough [21]), on curriculum concerns, and on discussion with staff in the Olympiad program. The dimensions are summarised in Table II.

The *content* dimension was used because the content knowledge covered by the ASOE exams has previously been raised as an equity concern. This is because, while students from all states in Australia may take the exam, there is not currently a national curriculum (although one is now being gradually implemented), and topics covered vary somewhat by state. Hence, while the exam content has not been previously considered a gender equity issue for the ASOE, it has been considered an equity issue on geographical grounds.

The *process* dimension is an obvious complement to the content dimension. As educators we are now used to considering not only what content knowledge we need to teach, but also what skills students need to learn. Listing acquired skills in statements of learning outcomes is now almost as commonplace as listing acquired knowledge. In addition, as the APhO and IPhO competitions frequently present students with questions in content areas that they are completely unfamiliar with, there is a strong focus in the Australian Science Olympiad (ASO) Physics Program on problem-solving techniques as well as content knowledge. Hence, testing a student's approach to problem solving is a significant aspect of the questions used in the selection process.

TABLE II. Dimensions used to describe the questions.

Dimension name	Description
Content	What students need to know: physics content knowledge required to answer the question, such as Newton's third law, conservation of momentum, or Ohm's law.
Process	What students need to be able to do: skills required to answer the question successfully, such as applying a principle, performing algebra, or interpreting graphs.
Presentation	How the information is presented: for example, whether it is primarily as words, equations or diagrams.
Context	Degree of abstraction from personal, physical experience.
Difficulty	Complexity of the question: for example, the number of concepts or steps required, or the number of spatial dimensions involved.

The *presentation* dimension is based on previous studies of gender differences in the preference for, and performance in, tasks involving interpreting words versus those that require the interpretation of a diagram or some other nontextual visualization [10,12,13]. It could be argued that the presentation dimension could be subsumed into the process dimension, as understanding a question is the first skill needed in attempting an answer. However, it was decided to retain presentation as a distinct dimension because it is a basic aspect of question design, and the visual appearance of a question may influence a student's initial reaction to a question, even before they have fully read the question. Many questions can be presented in a variety of ways (for example, using either graphical or textual descriptions of a process), and it may be that this choice influences how students perform on the question.

As already discussed, question *context* has been a concern of several previous studies [3,11]. However, we have not categorized questions by "gendered context" (as previously done by, e.g., Hazel *et al.* [2] and McCullough [21]), as in many cases it is difficult to determine how a question should be classified. In addition, recasting questions with apparently feminine or masculine contexts has not been shown to change gender gaps in any consistent way [17,24].

Instead, considering the work described by McCaskey *et al.* [36] and McCaskey and Elby [37], we looked at context in terms of "concrete" or "abstract": whether students were likely to be able to personally relate to the context used, and whether students may have had actual physical experience of the situation described in a question.

McCaskey *et al.* [36] and McCaskey and Elby [37] observed that female students are more likely to answer questions differently when asked to answer the question "as a scientist would" or "according to their own beliefs." When the context of a question clearly points to a real-world situation, or alternatively to a textbook-style physics question, the context alone may trigger this split in selecting an answer. We hypothesize that girls are more likely to answer a question framed in a typical "textbook" context using the principles of physics taught in textbooks and in the classroom. In other words, we contend that it is more likely that under such a scenario, girls will answer as they have been taught that a scientist would answer. In contrast, when a question context is realistic, females may be more likely to rely on their own beliefs and experience.

The dimension of *difficulty* gives an indication of the cognitive load placed on the student in correctly answering

the question. For example, a multistep question in which two equations must be applied requires more effort than a question in which only one equation is applied.

In summary, the content and process dimensions describe what students need to know or be able to do. The difficulty dimension gives an indication of the cognitive load placed on the student in correctly answering the question. The presentation and context dimensions look at the form in which the information is given and the question situated.

Within each dimension, a set of categories were constructed and refined using an iterative process. First, a substantial list of possible categories was constructed, with the intention that the list should cover all possible ways in which any given question could be classified within that dimension. The categories were defined as unambiguously and objectively as possible to ensure that categorization would be simple and reliable. The categories used in each dimension are described below.

This schema was then refined by sorting the questions into categories, noting that any given question could be classified into more than one category. At the end of this first iteration, empty categories were discarded, and categories with complete overlap within a single dimension were merged. At this stage a few categories were also added, where subtle differences between questions were not captured by the existing scheme.

The questions were then completely sorted a second time. Most questions fitted obviously into a given category: for example, a question requiring the application of conservation of momentum clearly fits into the "conservation of momentum" category. Where the classification was not so obvious, classification reliability was checked by having a second classifier categorize the question. In the 80 questions categorized, only two questions were shifted on the basis of disagreement between the two classifiers. More description of the different categories is given below and in Tables III–VII.

The data set was then explored to find patterns of gaps within each dimension.

V. ANALYSIS OF QUESTIONS BY TYPE

We now describe the patterns of gender differences (gaps) found when questions were sorted into categories within each dimension. Within each dimension every question was categorized as belonging to at least one category, and often to more than one category.

A. Content

When the questions were sorted by content, some patterns in gender gaps emerged, although these were complex, particularly for questions requiring knowledge of mechanics. Within the broad topic of mechanics there was a mix of gaps, including the only two questions with significant gaps favoring females, the questions with very large gaps favoring males, and a great variety in between. Some of the categories and the patterns of gaps that emerged are shown in Table III.

Note that most questions were classified as belonging to a single category within the content dimension, although a small number required content from two areas of mechanics: for example, the use of both the definition of kinetic energy and conservation of momentum. The "generic category" was used to classify questions where no specific physics content knowledge was required. The generic category was further divided into subcategories including order of magnitude estimates and algebra.

As shown in Table III, we found that gaps are generally large and in favor of males for questions involving kinematics and/or projectiles. Variations of FCI question 13 (where a boy throws a ball straight up, and the forces involved during the subsequent motion must be identified) were used twice (2007 Q1 and 2012 Q1). In both instances the gap was large (0.18 and 0.20) and significant ($\chi^2 = 21.7$ and 20.5, $\phi = 0.16$ and 0.19). Large gaps in male and female ability to answer this question have previously been observed [17,23]. These gaps may be explained by social or cultural effects, such as higher participation of males in physical activities involving throwing and catching from very young ages, leading to differences in the way spatial and kinetic information is processed by males and females [10,13]. We discuss this pair of questions in more detail in Sec. VI.

Gaps on questions involving Newton's laws are generally small, with a single exception being a question involving vertical acceleration (2007 Q5), on which the gap was large (0.17, $\chi^2 = 22.8$, $\phi = 0.17$) in favor of males. Questions on momentum have gaps from 0.10 in favor of females to 0.12 in favor of males. The larger male-favoring gaps are on questions where students are asked a straightforward question, such as "When two objects collide, which has the larger change in momentum?" The two female-favoring gaps, which are the *only* significant female-favoring gaps in the entire data set, occur when students need to apply conservation of momentum on the second of a pair of questions, without being explicitly told

TABLE III. Pattern of gender gaps within the *content* dimension.

Category	Gaps
Kinematics (including projectiles)	Generally large and favoring males
Newton's laws	Generally small
Conservation of momentum	Variable, including large gaps favoring each gender
Content outside of mechanics	Variable
Generic	Variable

to do so. These two pairs of questions are also discussed in more detail in Sec. VI.

Temizkan [38], who studied a group of year 10 students, noted that females are more likely than males to hold misconceptions described as impetus and gravity even after confounding factors such as school type and ability were accounted for. We see consistent large gaps in our data set on questions involving gravity, with the answer selections made by females indicating a bias towards the impetus-type misconception of "a force in the direction of motion."

Outside of mechanics, it was observed that questions dealing with other physics content had a range of gaps. However, as there were only a small number of these questions, it was not possible to confidently identify any patterns.

Questions classified by content as "generic" showed mostly small to average sized gaps. Questions requiring logical reasoning had small or zero gaps. Dimensional analysis questions had gaps around the overall average. The only questions of this type with notably large male-favoring gaps were those asking students to estimate orders of magnitude. The gap was particularly large (0.14, $\chi^2 = 16.8$, $\phi = 0.16$) for the estimation of the speed of a flying object (2013 Q2), where females exhibited a strong preference for an answer about an order of magnitude too low.

B. Process

Most questions were classified into a single category within the process dimension; the categories and pattern of gaps are summarized in Table IV.

As shown in Table IV, the process that students are required to perform to successfully answer a question appears to affect the gap between male and female facility. Questions requiring mathematical manipulations, such as performing algebra or interpreting equations, generally have small or zero gaps. This is also the case for questions where students need to apply logical reasoning or interpret a graph. However, when the questions contained a diagram that needed to be interpreted, the gaps were larger. The question with the largest overall gap in the entire data set (2012 Q9: gap 0.28, $\chi^2 = 40.5$, $\phi = 0.26$) required the interpretation of a diagram of two-dimensional projectile motion. This question is also discussed in Sec. VI.

It was interesting to note that when students needed to identify and/or compare forces, the gaps were generally small when only horizontal forces and motion were involved, but were large when vertical forces and motion were involved.

TABLE IV. Pattern of gender gaps within the *process* dimension. Note that "Identify forces" and "Compare forces" are subcategories within "Apply a physical principle".

Category	Gaps
Apply a physical principle	Varied
—Identify forces	Small to large in favor of males
—Compare forces	Small to large in favor of males
Perform algebra	Zero or small
Perform calculations	Zero or small
Interpret an equation	Zero or small
Apply logical reasoning	Zero or small
Interpret a diagram	Small to very large in favor of males

C. Presentation

Each question was classified as belonging to only a single category in the presentation dimension, except for a small number of questions that used both equations and numbers. The categories and emerging patterns are shown in Table V.

Gaps are generally average or larger than average where a diagram needs to be interpreted, particularly when the diagram contains information relating to two spatial dimensions. This is consistent with the findings of other research. Halpern *et al.* [10] and McBride [13] describe how males are better than females at processing visual information, particularly to do with spatial positions and motion.

Gaps are small on all questions where there is a large amount of reading to do. This result is not surprising, given the work by Nadeau and Quinn [39] demonstrating that males have shorter attention spans than females, and that females are more skilled at verbal communication. McBride [13] found that the more words used to convey instructions, the more likely it is that males will stop paying attention to the matter at hand. Hence, male students may not be reading the entire question, particularly when there are multiple lengthy answer statements to select from, or they may not be reading with sufficient care to extract any subtleties.

Gaps are generally small when equations and numbers are involved. This indicates that the females in this sample do not have greater difficulty in interpreting abstract information, such as equations, than the males in this

TABLE V. Pattern of gender gaps within the *presentation* dimension.

Category	Gaps
Primarily words: Short	Varied
Primarily words: Long	Zero to small
Significant diagram	Average to large in favor of males
Significant graph	Varied
Equation	Zero to small
Numerical data	Zero to small

sample. This is in contrast to suggestions by Gurian and Stevens [12] that males are more suited to dealing with abstractions. We note, however, that this sample may not be typical of the wider cohort of students, as it nominally represents only the top performing year 11 physics students.

Only a small number of questions required students to interpret graphs, and the gaps were mixed on these questions. In addition, the facility on one of these questions (2010 Q10) was extremely low (0.15 for both genders), so the data are limited. Other factors are likely to be important in determining the gaps on these questions.

D. Context

The context dimension was divided into five categories based on how students may be pushed to think about a question by the presentation of the question: relating it more to their own experience or intuition or common sense about real-world situations, or towards thinking about it as a "textbook" problem, to be answered using principles and techniques learned in the classroom. The categories used were real *and* possibly experienced, real *but not* personally experienced, contrived, "textbook," and none.

The "real and possibly experienced" category is for questions that involve something a student may have *physically* experienced, for example, throwing or catching a ball, or pushing a heavy box.

The "real but not personally experienced" category is for questions with situations that are realistically possible, but with which students are unlikely to have had a personal, physical experience. This category includes questions in which a vehicle hits another object. While some students may well have been in a vehicle collision, they were not themselves the vehicle experiencing the force. In such cases, some leap of imagination is required to identify with the objects involved.

The third category, "contrived," is for questions using obviously silly contexts, such as a collision between a chicken and a paper plane in midair. This is a separate category to the previous one, because it invokes typical contrived textbook scenarios, rather than arguably real-life situations, and hence may push students to answer questions using physics principles rather than their "common sense" or genuine beliefs.

"Textbook" contexts are those involving blocks on frictionless surfaces, ideal pulleys, or similar. Questions in which students are asked to manipulate an equation with no relevant context are categorised as "none."

Of the five dimensions considered, the clearest pattern emerges from classifying the questions by context. This is summarized in Table VI.

We can consider the contextual types as a spectrum from most concrete to most abstract, with "real world and experienced" at the most concrete end of the spectrum through "real world but not experienced." then "contrived." "textbook," and, finally, "none" at the most abstract end. We see that, with some exceptions, gaps between male and female facility decrease from large and in favor of males at the more concrete end of this spectrum, down to near zero at the more abstract end. Our findings are in contrast with those of Rennie and Parker [20] who found that females prefer concrete over abstract contexts. Hazel *et al.* [2], on the other hand, found that both males and females performed better on questions with no context than on questions with context, with a larger relative gap on questions with context. This is similar to our findings.

TABLE VI. Pattern of gender gaps within the *content* dimension.

Category	Gaps
Real and possibly experienced	Generally large gaps in favor of males
Real but not personally experienced	Varied gaps, but mostly small to large in favor of males
Contrived	Small to average gaps in favor of males
Textbook	Varied, but generally small gaps
None	Zero to small gaps

We find that gaps are generally large on questions where we expect at least some students to have had personal, physical experience of the situation described, such as throwing and catching objects. It is likely that the difference in gaps for real-world questions where students may have experienced the situation and real-world situations where they have not is due to the generally more extensive personal experience of boys compared to girls in physical activities involving, throwing, pushing, and lifting.

An interesting exception to this general observation involves questions similar to FCI question 26, relating to the motion of a box being pushed along the floor. We find *small* gaps on these questions, and it may be that for these questions personal experience is misleading.

E. Difficulty

The categories used to describe the difficulty dimension were the number of steps required, the number of concepts required, and the number of spatial dimensions to be considered. Most questions were classified three ways in this dimension: needing one, two, or more steps; needing one, two, or more concepts; and needing one, two, or more spatial dimensions. Some questions (such as algebraic manipulations) were only categorized by the number of steps required. The results for this categorization are shown in Table VII.

No clear relationship between the number of concepts required and the gap between male and female facility was evident in the data. There does appear to be a small increase

TABLE VII. Pattern of gender gaps within the *difficulty* dimension.

Category	Gaps
Steps: One step	Varied
Steps: More than one step	Varied, but slightly larger and in favor of males than for the "one step" category
Concepts: One concept	Varied
Concepts: More than one concept	Varied
Dimensions: One spatial dimension	Varied, mostly small
Dimensions: More than one spatial dimension	Large to very large in favor of males

in average gap when two or more steps are required compared to only a single step, although notable exceptions to this are the two questions that showed a large bias in favor of females (which required at least three substantial steps). Gaps tended to be larger where individual steps were small or of a similar nature.

The dimensionality of a problem has the clearest impact on gender gap in the difficulty category. Questions where students needed to consider the problem in two dimensions—particularly when there was a relationship between horizontal and vertical forces, velocities, or positions—had large gaps in favor of males. Examples of this are the two-dimensional projectile motion question (discussed in detail later), which had the largest overall gap, and a question where force magnitudes needed to be compared with the forces acting vertically, horizontally, and at an angle (2014 Q6).

Docktor and Heller [15] and Madsen *et al.* [14] found in their studies that the largest gender gaps were associated with FCI questions 14 (identifying the path of an object released by an airplane) and 23 (identifying the path of a rocket once the thrust is removed), while Dietz *et al.* [16] found large gender gaps for FCI question 23, together with FCI questions 6 (identifying the path of a ball leaving a channel) and 12 (identifying the path of a cannonball fired horizontally from a cliff). All of these questions require a mental model of motion in two dimensions, as is the case for the ASOE question with the largest gap (2012 Q9). The picture that emerges is consistent with the literature [12,13]. It appears that females find it harder to create mental models of motion in two dimensions, whether because of underlying differences in brain structure and function or from lack of relevant early experiences, or an interaction between the two.

In summary, as Madsen *et al.* [14] point out, the overall gender gap is likely to be a combination of many small factors. Based on our analysis, important factors include question *content*, the *presentation* of the information in the question, and the *context* (defined as the degree of abstraction from experience).

VI. QUESTIONS OF PARTICULAR INTEREST

We will discuss four questions of particular interest here. First, we will discuss the two questions (2007 Q1 and 2012 Q1) similar to FCI question 13 regarding the forces acting on a projectile in one dimension, which has previously been identified as showing a large gender gap in facility. Second, we will discuss the two-dimensional projectile motion question (2012 Q9), which showed the largest gap of all questions. Third, we will discuss the paired kinetic energy and momentum questions (2010 Q1 and Q2 and 2011 Q5 and Q6), which show a male bias on the first of the pair, and the only large female biases in the entire data set on the second of the pair.

A. Force on a projectile in one dimension

Question 13 of the 1995 version of the FCI [5] asks students to identify the forces acting on a steel ball, which has been thrown directly upwards, after it has left a boy's hand. Popp *et al.* [23] and Bates *et al.* [17] have previously identified this question as having a bias favoring males. McCullough [21] used a revised version of this question, in which a girl throws a teddy bear straight upwards, but the question is otherwise unchanged. McCullough's version was used as Q1 of the 2007 ASOE, and a variation in which "Maddie throws her phone straight up" was used as Q1 in 2012.

In 2007, the male facility on this question was 0.389 and the female facility was 0.212; hence, the gap was 0.177. In 2012, the male facility on this question was 0.527 and the female facility was 0.328; hence, the gap was 0.199. Past exam papers were available on the internet for students and teachers to download prior to 2012, so training may account for the increase in facilities between 2007 and 2012. However, despite the increased facilities, the large gap is maintained and is statistically significant at much better than the 1% level ($\chi^2 = 21.7$ and 20.5, $\phi = 0.16$ and 0.19, respectively).

To understand how this gap arises, we need to look at the distribution of answers chosen by males and females. Figure 3 shows the item response curves for the 2012 version of this question for males and females. The item response curves for the 2007 version of this question are very similar.

Morris *et al.* [32] note from the item response curves for FCI question 13 for their sample that the two main distractors are answers B and C, with lower-achieving students preferring answer B and midachieving students preferring answer C.

We can see from Fig. 3 that females are more likely than males to choose distractors B and C, particularly distractor C for mid-scoring females, much as Morris *et al.* [32] observed. Distractors B and C both include a steadily decreasing upwards force on the thrown object during its ascent. Distractor B includes a varying gravitational force,

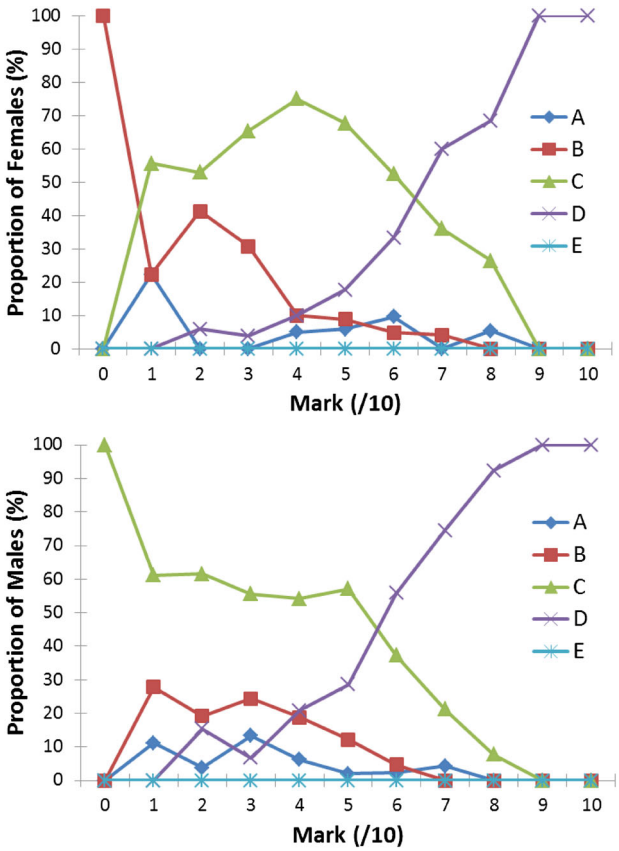


FIG. 3. Item response curves for the 2012 version of the force on a projectile in vertical motion question (FCI question 13), for females (top) and males (bottom). The horizontal axis shows the students' scores; the vertical axis is the proportion of students with that score choosing each of the five possible answers.

while distractor C has a constant gravitational force. These two distractors were chosen not only by girls with low overall scores, but by a significant number of females who scored at least 7/10 in the MCQ section of the ASOE. In contrast, these distractors were chosen only by low- to mid-scoring males.

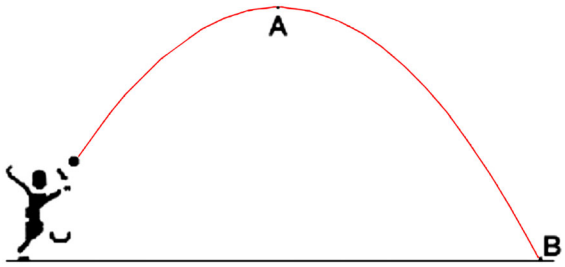
This choice of distractors is consistent with the results of Temizkan [38], who found that girls are more likely to hold the "impetus" misconception, as well as misconceptions about gravity. This misconception seems to be held by lower-scoring students of both genders, and by high-scoring females students, but *not* by high-scoring male students. In our categorization scheme this question was classified by content as "forces—gravity" and by context as "real, not experienced." Both of these types show higher than average gaps, and hence the combination of real-world context and the projectile motion content appear to combine to give a large gap on this question. It is possible that asking the question with the same content, but with a less real-world context would reduce the gap.

B. Projectile motion in two dimensions

The question with the largest overall gap in the ASOE set (2012 Q9) asks students to compare the speed and acceleration of a ball at two different points in a two-dimensional trajectory. The question is shown in Fig. 4.

The average male facility on this question was 0.673, compared to an average female facility of 0.397. This gives a gap in facility of 0.276 in favor of males, which is significant at far better than the 1% level ($\chi^2 = 40.5$,

9. A ball is thrown into the air and it moves in the path shown below. Ignore air resistance in this question.



At position A the ball is at the highest point in its path, position B is just before it hits the ground. Which of the following statements is true?

A. The speed of the ball at A is zero and the acceleration of the ball at B is the same as at A.

B. The speed of the ball at A is the same as the speed at B and the acceleration at B is higher than at A.

C. The speed at A is lower than the speed at B and the acceleration at A is higher than the acceleration at B.

D. The speed at A is lower than the speed at B and the acceleration at A is the same as the acceleration at B.

E. The speed at A is higher than the speed at B and the acceleration at A is the same as the acceleration at B.

FIG. 4. Question 9 of the 2012 ASOE. This question showed the largest gap in facility in the entire ASOE data set.

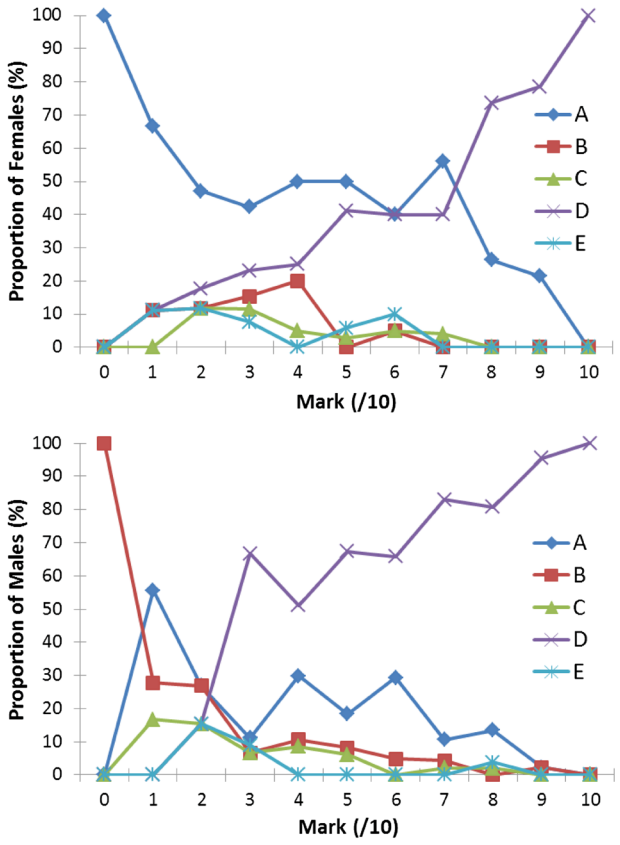


FIG. 5. Item response curves for question 9 of the 2012 ASOE, Projectile motion in two dimensions, for females (top) and males (bottom).

$\phi = 0.26$). Item response curves for this question are shown in Fig. 5.

Figure 5 clearly shows that the gap is due to the female students' preference for answer A, which incorrectly identifies the speed of the ball at point A as zero. This distractor is only a slightly more popular choice than the other incorrect answers for male students, but it is the overwhelming choice for female students, persisting as the most common choice for female students scoring up to and including 7/10 overall. The score at which the correct answer is more commonly chosen than any distractor (the "crossover" point) occurs at around 2/10 for male students, which is much lower than the average mark of 5.8/10 for male students, compared to a crossover of more than 7/10 for female students, which is higher than the average mark of 5.2/10 for female students. In other words, the female preference for "speed is zero at the apex" is a significant, consistent, and dominant misconception, indicating confusion between pure vertical (one-dimensional) motion and combined vertical and horizontal (two-dimensional) motion.

This question was classified by content as both "forces—gravity" and "kinematics." In the context dimension it was classified as "real but not personally experienced" and by presentation and process as "diagram," as it is necessary to interpret the diagram to correctly answer the question. It was also classified as two dimensional in the difficulty category. Hence, in all of the five dimensions this question was classified as of a type that was shown to have larger than average gaps in facility. All of these male-favoring factors have combined in this single question to give an extremely large gender bias in favor of males.

To try to understand why at least some of these factors result in such a large bias, we turn to the neuroscience and psychology literature. Halpern *et al.* [10] review the literature on gender differences in cognition, in the context of success in science and mathematics. They state that while females are more likely to excel in verbal activities, males outperform females on most measures of visual-spatial ability. In addition, as males are more variable in visual-spatial ability, there are more males at the very high ability end. These findings are consistent within the literature: McBride [13] notes that males are more focused on spatial aspects of a situation such as movement; and Maitland *et al.* [40] note that, while differences decrease with age, males outperform females on spatial manipulations across all age groups. Hence, it is not surprising that this question, which requires the interpretation of a diagram representing movement, together with a mental model for a projectile moving in two dimensions, is done poorly by female students.

C. Kinetic energy and conservation of momentum

In the entire set of 80 ASOE questions, only two showed a large gap in favor of females. These questions (2010 Q2 and 2011 Q6) were essentially the same question, with different context, and required students to apply conservation of momentum. In both cases, the questions were the second in a pair dealing with a collision between two airborne objects: in 2010 these were "a large, hefty chicken" and "a small, light quail"; and in 2011 they were "a chunky tree frog" and "a small speedy fly." The 2010 version of this pair of questions is shown in Fig. 6. The 2011 version is similar, but with answers A and B swapped in both questions, and no numerical speed included in answer D in the second question of the pair. Here, we discuss the gender differences in both questions of the pair.

The facilities on 2010 Q1 were 0.709 and 0.827 for females and males, respectively, giving a gap of 0.118 in favor of males, which is statistically significant at better than the 1% level ($\chi^2 = 10.4$, $\phi = 0.13$). The most popular incorrect answer for both genders was C ("same kinetic energy means same speed"), but D ("kinetic energy and speed not related") and E ("direction matters") were also common choices for low- to mid-scoring females, as shown in the item response curves in Fig. 7.

Figure 7 shows that the gender gap on this question arises from the different answer choices of low- to mid-scoring students. High-scoring students (7/10 and above) all answered this question correctly. The bias towards

1. A large hefty chicken and a small light quail, both flying in midair, have the same kinetic energy. Which of the following statements is true?

   A. The chicken has a greater speed than the quail.

   B. The quail has a greater speed than the chicken.

   C. The chicken and the quail have the same speed.

   D. Nothing can be inferred, the kinetic energy has nothing to do with the speed.

   E. The direction of the quail and the chicken must be taken into account before a decision can be made.

2. The large hefty chicken and the small light quail, still both flying in midair with the same kinetic energy, fly directly at each other and collide head-on due to a joint navigational error, and briefly become one fowl object. Which of the following statements is true in the instant after the collision?

   A. The object moves in the same direction as the chicken's original motion.

   B. The object moves in the same direction as the quail's original motion.

   C. The object stops dead in the air.

   D. As soon as they collide they move downward with velocity 9.8 $\mathrm{m\,s^{-1}}$.

   E. It depends on how hard the chicken hits the quail.

FIG. 6. Questions 1 and 2 from the 2010 ASOE. Question 1 showed a large gap in favor of males, while question 2 showed a large gaps in favor of females.

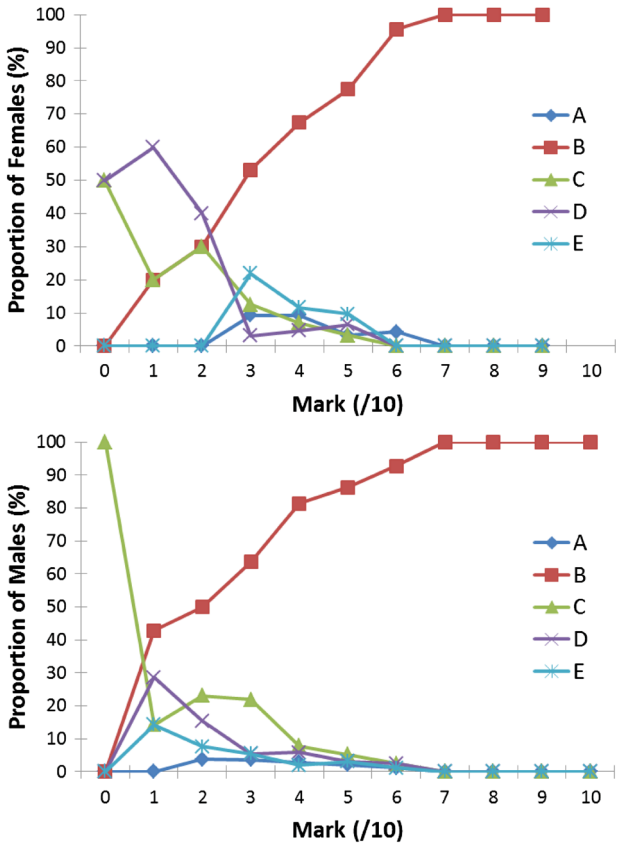


FIG. 7. Item response curves for question 1 of the 2010 ASOE, *kinetic energy and speed*, for females (top) and males (bottom). No females achieved a score of 10/10 in 2010.

answer D ("kinetic energy and speed not related") appears mainly among the low-scoring females, while the mid-scoring females show a bias towards E ("direction matters"), which indicates a more complex, though still incorrect, model of energy. Alternatively, this response may be linked to those mid-scoring females experiencing a lack of confidence, and preferring a less absolute answer.

The facilities were slightly smaller in 2011; 0.676 for females and 0.800 for males, gap of 0.124 in favor of males ($\chi^2 = 12.5$, $\phi = 0.13$, so significant at better than 1%). The gap was again due to differences in answer selection for low- and mid-scoring students, with the crossover from incorrect to correct again occurring at lower scores for males.

The large gap for this question can be understood in much the same way as that for the first question discussed in this section (FCI question 13). This question deals with objects (projectiles) moving through space, and males appear to have better mental models of projectile motion. In addition, female students in this cohort appear less likely to either have the basic content knowledge (that kinetic energy is a function of speed and mass) or are not able to apply it in this context. The change of context from "chicken and quail" (where the masses are different, but same order of magnitude) to "frog and fly" (where the masses are orders of magnitudes different) appears to make little if any difference to both genders.

The more interesting question of the pair is the following one, where students are asked what happens when the objects collide. Females were much more likely to answer this question correctly compared to males, in contrast with the preceding (but related) question.

The facilities on 2010 Q2 were 0.442 for females and 0.344 for males, respectively, giving a gap of 0.098 in favor of females, which is statistically significant at the 5% level ($\chi^2 = 5.1$, $\phi = 0.091$). The facilities for 2011 Q6 were much higher (0.725 for females and 0.628 for males), but the gap remains the same (0.098 in favor of females, $\chi^2 = 6.2$, $\phi = 0.0.093$, significant at the 5% level). The increased facility from 2010 to 2011 is likely to be due to training, as past papers were available online.

The item response curves, Fig. 8, give us some insight into how this gap arises.

The item response curves indicate that answers C ("combined object stops dead") and D ("combined object immediately moves downwards") are the most popular distractors, and that they are persistent even for high-scoring males. Females discard these choices in favor of the correct answer at lower scores. In 2011, the same pattern was observed, but with a lower crossover point for both genders. In almost every other question involving projectile motion there is a large gap in favor of males. Hence, it

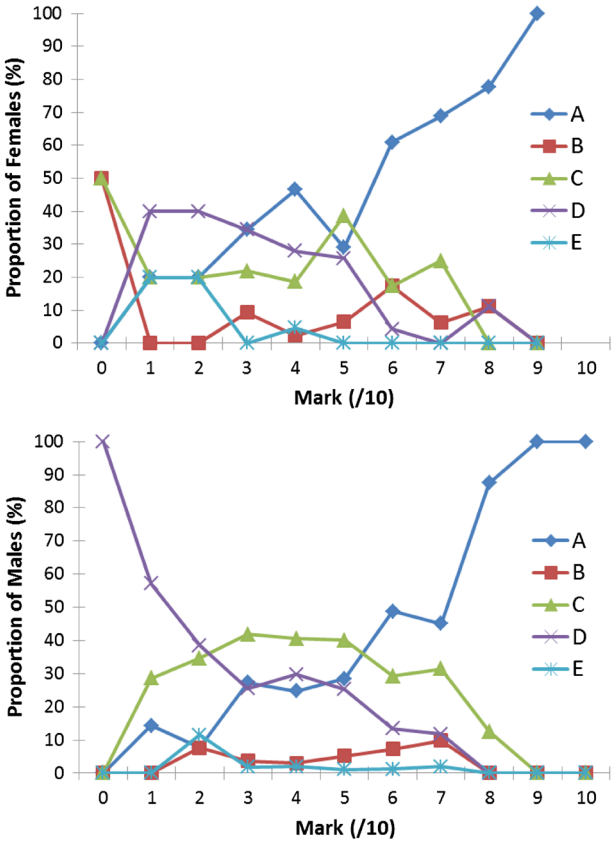


FIG. 8. Item response curves for question 2 of the 2010 ASOE, *collision of two objects with same kinetic energy*, for females (top) and males (bottom).

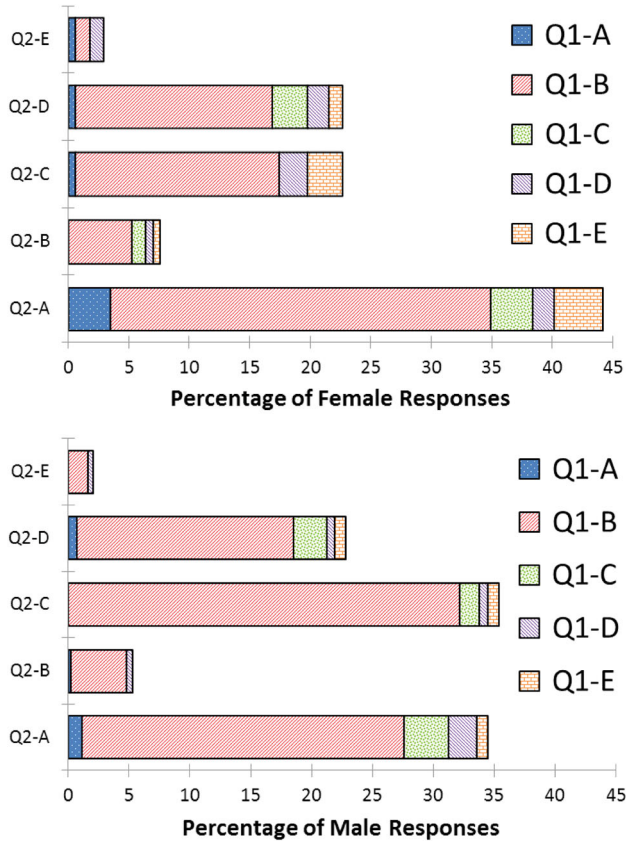


FIG. 9. Female (top) and male (bottom) answer choices for question 2 of the 2010 ASOE, *collision of two objects with same kinetic energy*. The colors or shading within each horizontal bar indicate the choices made by those students on the previous question (2010 Q1, *kinetic of two objects*). The correct choice pair is Q1-B (red or striped upwards) and Q2-A (lowest horizontal bar).

appears that males have better mental models of projectile motion, possibly due to more exposure from a very young age to activities such as throwing and catching balls, for example, while playing school sports.

It is interesting to consider the reasoning required to answer the question correctly. To answer both questions correctly, a student needs to use the definition of kinetic energy ($K = \frac{1}{2}mv^2$) at the first question, and then the relationship between kinetic energy and momentum ($p = \sqrt{2mK}$) *and* conservation of momentum at the second question. Figure 9 shows how students progressed from 2010 Q1 to 2010 Q2.

We can see from both Figs. 8 and 9 that answers C and D are more common distractors for males than they are for females. Figure 9 shows that many males select one of these incorrect answers at Q2, despite having answered Q1 correctly (the "incorrect" male Q2-C and Q2-D bars show a large component of "correct" Q1-B answers).

Figure 9 also shows that a significant number of females are selecting the correct answer at 2010 Q2, after *not* correctly identifying the relationship between kinetic energy and speed at Q1 (the "correct" female Q2-A bar has contributions from many incorrect Q1 answer selections). These students are not following the chain of reasoning described above, which is required to correctly answer Q2. Instead, they are relying on something else, most likely their "intuition." If you imagine watching the collision, it makes sense that when one moving object collides with a very much smaller moving object the combined object continues on with much the same velocity as that of the larger object.

In order to gain some insight into how students may have arrived at their answers, this pair of questions was given to a group of first-year university students who were studying physics, composed of 16 males and four females. The students were asked to attempt the pair of questions, and make notes on how they chose their answers. They were then engaged in discussion about how they approached the questions, whether they used visualization, equations, or something else. While these numbers are not large enough for statistical analysis, particularly to look at gender differences, it did give some insight into possible ways students are conceptualizing the questions.

Most of the students in this group answered the first question in the pair correctly, and the second incorrectly. Those who answered the second question incorrectly

generally chose answer C or D on the grounds that the kinetic energies would "cancel out," as they were equal but in opposite directions. Several had written next to the questions "$KE_1 = -KE_2$", "$KE = 0$", or similar statements. Of those students who answered correctly, all but one described imagining the collision, and the large object carrying the small object along with it. Only one student used the sequence of equations described above. Two students initially selected the correct answer, then changed to answer C on the grounds that when they "thought about it more carefully and used the equations properly" and canceled kinetic energies, the correct answer must be C.

It appears that female students, who are more likely to answer the first question incorrectly, may be more likely to select the "intuitive" response: that a large moving object colliding with a small object will tend to carry the small object along with it. Males, who are more likely to correctly apply the definition of kinetic energy at the first question, may then go on to misapply the concept of kinetic energy at the second question and select the incorrect answer. The strong cuing in the questions to consider kinetic energy makes this more likely; kinetic energy is mentioned in both questions, while momentum is not mentioned at all. It seems that only very high ability students make the link to momentum conservation on their own—those few of either gender that answer both questions correctly.

To determine whether this is indeed what is happening in these questions, further investigation with larger numbers of students is required.

VII. IMPLICATIONS

From the outset, remember that the students in this study are selected by their teachers as being of high ability or "good at physics," and hence may not be representative of the larger cohort. Our results indicate that the females in this cohort perform just as well on reasoning and maths questions as the males. If we take their performance on reasoning questions as a proxy of ability (see, e.g., Refs. [41–43]), then the overall ability of the females in this sample is similar to that of the males. However, there is a statistically significant gap in male and female overall performance on the ASOE, including the MCQ section, which is consistent over the eight years of this data set, and had been observed in earlier years [1].

It appears from our results either that the females in this cohort have less physics content and procedural knowledge (or the ability to apply that knowledge) or that many of the questions are biased in some way. It seems likely that both effects are contributing to the large gaps that we observe. There are particular content areas, including projectile motion, where female students consistently perform poorly compared to male students. This occurs regardless of the question context, or the way in which the information in the question is presented. We have also observed that the way in which a question is presented can affect the gap: questions that are situated in a real-world context, particularly ones that males are more likely to have experienced, are generally done better by males.

There is significant literature that claims that male brains are better at modeling objects moving through space, and that this is due to structural differences in males' and females' brains. However, as Halpern *et al.* [10] states, "Experience alters brain structures and functioning, so causal statements about brain differences and success in math and science are circular."

Female students could be helped to develop better mental models of how objects move through space by giving them more of the opportunities that boys are given in early childhood. This includes active physical activities such as playing ball sports and engaging with toys typically listed in the "boys"' section of toy catalogs, such as gliders, water rockets, and projectile launchers. In Australia, there is still a tendency for girls and boys to be treated differently from a very young age: this includes being given different toys, being presented with different role models, and being dressed differently. Small girls are more likely to be dressed in clothing that inhibits running, jumping, and engaging in vigorous physical activity than is the case for small boys. It seems likely that at least some of the gender gap that we see in physics tests at year 11 can be traced back to experiences in early childhood.

By the time students reach year 11, however, our opportunity to broaden their experience is limited, and the societal changes required to ensure that girls are given the same opportunities as boys are likely to take another generation (or longer). Instead, we need to think about what it is that we are testing for, and how we can most effectively do so. We need to consider what it means to be "good at physics." Obviously, an understanding of projectile motion and the ability to interpret two- or three-dimensional diagrams will still be a part of this, particularly as physics forms the basis of other disciplines, such as engineering, where these skills are central.

However, being good at physics should also include being "good at thinking"; that is, to be able to reason logically from evidence, and to think critically about explanations. The ability to interpret lengthy text-based descriptions or instructions is also necessary for working physicists, and other people who use physics. These latter skills, which are generally more highly developed in girls, are also part of being "good at physics" in the real world, although they are often not emphasized in the classroom or on test papers. While typical physics tests contain MCQs or short written questions on projectiles, the use of scientific reasoning questions, for example, is still fairly unusual. If we want our assessment to reflect the broad range of content knowledge, skills, and ways of thinking that make for a "good physicist," then we need to broaden what, and how, we test.

The IPhO examinations consist of three very difficult theoretical problems in one five-hour examination, and one

or two experimental problems to be solved in a second five-hour examination. The exams require students to demonstrate not only depth and breadth of content knowledge, but also highly developed problem-solving skills. Students are often faced with questions in an unfamiliar content area, and need to apply a range of skills including reading, visualization, mathematics, and reasoning. In the laboratory examination students need to demonstrate practical, hands-on skills as well as problem-solving and data analysis skills. If we consider that these skills are a significant part of what makes a "good physicist," then the IPhO examinations are testing for students who are "good at physics."

In the case of the ASOE, the aim is to ultimately select a team to compete at the IPhO competition. At present, and for many years, girls have been underrepresented in most of the teams (from Australia and elsewhere) at the international competition. In fact, Australia (on average) has a significantly higher number of female participants than most countries.

Australian Science Innovations has a strong motivation to ensure that all students, regardless of gender, can access the selection process to choose the best possible team. However, it is not logistically practicable to use several hours of paper-based and practical examinations at the first stage of the selection process. Hence, the ASOE is a relatively blunt instrument that seeks to identify students with enough physics content knowledge, along with sufficiently well-developed mathematical and reasoning skills, to give them the potential to become "good at physics" as measured by performance in the IPhO examinations the following year. It is thus highly undesirable to "filter out" students too early in the selection process, particularly based on factors such as gender.

We have identified patterns in gaps by question types, particularly in the content, context (real to abstract), and presentation dimensions. Based on these findings, it should now be possible to design an exam that reduces the gaps. We should be able to reduce, even if not eliminate, the gaps in content areas that female students find difficult, by changing the way in which information is presented and setting the question in a context less likely to favor typical male experiences. This will reduce selection bias in the exam, giving more females the opportunity to participate in the team selection process. These ideas can be tested in future ASOEs.

While we have focused on a particular examination, with a particular purpose, the findings from this work are relevant to the use of MCQs in physics testing more generally. MCQs are commonly used for both diagnostic testing and assessment of learning. Instructors who use MCQs need to be aware of possible gender biases in the questions that they use, and consider factors such as context and information presentation when writing MCQs.

There is still much work to be done. Further research, including interviews with students, is needed to investigate whether different genders are modeling similar situations in different ways. There is significant scope for further research to be done on these data and results. There is a wealth of information contained in the written question section of the ASOE, and a detailed analysis of these data may give further insight into gender differences in approach to answering physics questions.

We note that the gender gaps we have observed in the ASOE are, in many cases, similar to those observed in studies carried out in other countries [14], including the United States [15,16], the United Kingdom, and Ireland [3,17,24], which are culturally fairly similar to Australia. It would be interesting to compare the assessments used, and the typical gender gaps on assessment, in societies that are culturally more different.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Australian Science Innovations for access to the deidentified data used and their support, and to Darren Goossens for his help with LaTeX. This research was conducted with approval from the UNSW Canberra Human Research Ethics Advisory Panel, Approval No. A-14-39.

[1] K. Wilson, N. Kueter, G. Dennis, M. Verdon, and A. Nulsen, Addressing gender disparity in the physics national qualifying exam for the Australian Science Olympiads, Teach. Sci. **53**, 24 (2007).

[2] E. Hazel, P. Logan, and P. Gallagher, Equitable assessment of students in physics: Importance of gender and language background, Int. J. Sci. Educ. **19**, 381 (1997).

[3] S. Close and G. Shiel, Gender and PISA mathematics: Irish results in context, Eur. Educ. Res. J. **8**, 20 (2009).

[4] D. Hestenes, M. Wells, and G. Swackhamer, Force concept inventory, Phys. Teach. **30**, 141 (1992).

[5] I. Halloun, R. Hake, E. Mosca, and D. Hestenes, *Force Concept Inventory* (1995), http://modeling.asu.edu/R&E/Research.html.

[6] https://www.asi.edu.au/programs/australian-science-olympiads/past-exams/.

[7] E. E. Maccoby and C. N. Jacklin, *The Psychology of Sex Differences* (Stanford University Press, Stanford, CA, 1974).

[8] Eurydice, Gender differences in educational outcomes: Study on the measures taken and the current situation in Europe, http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/120en.pdf.
[9] C. Postles, Girls' learning: Investigating the classroom practices that promote girls' learning, edited by K. Moore, A. Reilly, and R. Naylor, https://plan-international.org/girls-learning-investigating-classroom-practices-promote-girls-learning.
[10] D. F. Halpern, C. P. Benbow, D. C. Geary, R. C. Gur, J. S. Hyde, and M. Ann Gernsbacher, The science of sex differences in science and mathematics, Psychol. Sci. Publ. Interest **8**, 1 (2007).
[11] J. S. Hyde, E. Fennema, and S. J. Lamon, Gender differences in mathematics performance: A meta-analysis, Psychol. Bull. **107**, 139 (1990).
[12] M. Gurian and K. Stevens, With boys and girls in mind, Educ. Leader. **62**, 21 (2004).
[13] W. McBride, *Teaching to Gender Differences: Boys Will be Boys and Girls Will be Girls* (World Books, Nashville, 2009).
[14] A. Madsen, S. B. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).
[15] J. Docktor and K. Heller, Gender differences in both Force Concept Inventory and introductory physics performance, AIP Conf. Ser. **1064**, 15 (2008).
[16] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the Force Concept Inventory?, AIP Conf. Ser. **1413**, 171 (2012).
[17] 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).
[18] L. J. Rennie and L. H. Parker, Assessment in physics: Further exploration of the implications of item context, Aust. Sci. Teachers J. **39**, 28 (1993).
[19] L. L Rennie and L. H Parker, Placing physics problems in real-life context: Students' reactions and performance, Aust. Sci. Teachers J. **42**, 55 (1996).
[20] 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).
[21] L. McCullough, Gender, context, and physics assessment, J. Int. Wom. Stud. **5**, 20 (2004).
[22] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).
[23] S. E. Osborn Popp, D. E. Meltzer, and C. Megowan-Romanowicz, in Proceedings of the Annual Meeting of the American Educational Research Association, New Orleans, 2011), http://modeling.asu.edu/R&E/FCI%20DIFanalysisAERA2011.doc.
[24] N. Walet and M. Birch, An update on gender differences on the FCI, http://www.iop.org/activity/groups/subject/hed/calendar/info/file_62073.pdf (unpublished).
[25] J. Harding, Sex differences in examination performance at 16+, Phys. Educ. **14**, 280 (1979).
[26] R. J. L. Murphy, Sex differences in objective test performance, Br. J. Educ. Psychol. **52**, 213 (1982).
[27] G. Ben-Shakar and Y. Sinai, Gender differences in multiple-choice tests: The role of differential guessing tendencies, J. Educ. Measure. **28**, 23 (1991).
[28] A. Tolmie and C. Howe, Gender and dialogue in secondary school physics, Gender Educ. **5**, 191 (1993).
[29] C. V Gipps and P. Murphy, *A Fair Test?: Assessment, Achievement, and Equity* (Open University Press, Philadelphia, PA, 1994).
[30] C. T. Richardson and B. W. O'Shea, Assessing gender differences in response system questions for an introductory physics course, Am. J. Phys. **81**, 231 (2013).
[31] T. Lyons, J. Kennedy, and F. Quinn, The continuing decline of science and mathematics enrolments in Australian high schools, Teach. Sci. **60**, 34 (2014).
[32] G. A. Morris, N. Harshman, L. Branum-Martin, E. Mazur, T. Mzoughi, and S. D. Baker, An item response curves analysis of the Force Concept Inventory, Am. J. Phys. **80**, 825 (2012).
[33] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Lawrence Erlbaum Assoc., Hillsdale, NJ, 1988).
[34] R. Coe, It's the Effect Size, Stupid: what effect size is and why it is important, in *Proceedings of the Annual Conference of the British Educational Research Association University of Exeter, 2002* (2002), http://www.leeds.ac.uk/educol/documents/00002182.htm.
[35] T. Baguley, Standardized or simple effect size: What should be reported?, Br. J. Psychol. **100**, 603 (2009).
[36] T. L. McCaskey, M. H. Dancy, and A. Elby, Effects on assessment caused by splits between belief and understanding, AIP Conf. Proc. **720**, 37 (2004).
[37] T. McCaskey and A. Elby, Belief vs. 'understanding': Why do students 'split' on the FCI? (unpublished).
[38] D. Temizkan, Master's thesis, Middle East Technical University, 2003, http://etd.lib.metu.edu.tr/upload/1207588/index.pdf.
[39] K. Nadeau and P. Quinn, *Gender Issues and AD/HD: Research, Diagnosis and Treatment* (Advantage Books, Washington, DC, 2002).
[40] S. B. Maitland, R. C. Intrieri, W. K. Schaie, and S. L. Willis, Gender differences and changes in cognitive abilities across the adult life span, Aging Neuropsychol. Cognit. **7**, 32 (2000).
[41] V. P. Coletta, J. A. Phillips, and J. J. Steinert, Interpreting Force Concept Inventory scores: Normalized gain and SAT scores, Phys. Rev. ST Phys. Educ. Res. **3**, 010106 (2007).
[42] P. Nieminen, A. Savinainen, and J. Viiri, Gender differences in learning of the concept of force, representational consistency, and scientific reasoning, Int. J. Sci. Math. Educ. **11**, 1137 (2013).
[43] L. Ding, Verification of causal influences of reasoning skills and epistemology on physics conceptual learning, Phys. Rev. ST Phys. Educ. Res. **10**, 023101 (2014).