PERSISTENT GENDER GAPS IN FIRST-Year PHYSICS ASSESSMENT QUESTIONS

David J. Low[a], Kate F. Wilson[b]

Presenting Author: David Low (d.low@adfa.edu.au)
[a]School of Physical, Environmental and Mathematical Sciences, UNSW Canberra, Canberra ACT 2610, Australia
[b]School of Engineering and Information Technology, UNSW Canberra, Canberra ACT 2610, Australia

KEYWORDS: gender gaps, physics, first-year

ABSTRACT
In a review of six years of first-year physics assessment results, we have identified a number of individual questions which display a persistent gender gap in performance: the facility displayed by males on these questions is consistently higher than that displayed by females. We have looked for patterns in student answers to these questions by gender and overall performance, and identified reasons why these questions might be more challenging for females. Our results indicate that the presentation and wording of questions can result in undesirable cueing, particularly when words and concepts such as 'positive' need to be uncoupled from their near-synonyms (such as 'upwards' and 'increasing') to be correctly applied in a physics context. The observations presented here suggest that this issue is more significant for females than it is for males. Since individual questions may contribute disproportionately to any overall gender gap in assessment, we encourage other educators to review their own historical records and assessment questions.

Proceedings of the Australian Conference on Science and Mathematics Education, Curtin University, Sept 30th to Oct 1st, 2015, pages 118-124, ISBN Number 978-0-9871834-4-6.

INTRODUCTION
BACKGROUND
Studies of gender differences as related to ability and achievement are widespread in the educational research literature, particularly since the work of Maccoby & Jacklin (1974). The main questions remain open: are these differences innate or acquired? Do they apply regardless of context? Can they be overcome by experience and/or education? Attempts to answer those questions often result in answers that vary depending on the study, and the particular field of enquiry, but continue to draw high-level attention, as evidenced by the reports of Eurydice (2010) and Postles (2013).

Three decades of work on gender differences in the sciences and mathematics are reviewed by Halpern, Benbow, Geary, Gur, Shibley Hyde & Gernsbacher (2007). There is evidence, albeit with many caveats, that females tend to excel in 'verbal' activities; while males demonstrate better performance in 'visual-spatial' tasks, and in tasks that require knowledge to be applied in a 'real world' context. In physics, the review by Madsen, McKagan & Sayre (2013) concluded that observed gaps in standardised tests are likely to be the result of a combination of many small factors (including socio-cultural effects such as the stereotype threat, self-efficacy, and teacher attitudes), and suggested that isolated explanations need to be treated with caution due to a lack of repeatability (as seen by McCullough 2004, for example). One common theme, however, appears to be the underperformance of females relative to males on questions that involve vertical and/or two-dimensional motion (see e.g. Docktor & Heller 2008; Dietz, Pearson, Semak & Willis 2012; and Bates, Donnelly, MacPhee, Sands, Birch & Walet 2013). Meltzer (2005), when discussing the role that the representational format of physics problems has in student performance, noted that females tended to underperform in questions that involved graphical (rather than verbal, diagrammatic, or mathematical) representations.

Multiple-choice questions (MCQs) have also been identified by a number of studies as being problematic for females. While the underlying reasons remain unclear and open to question, most suggestions in the literature centre on a male preference for taking a 'strategic' elimination-based black-and-white approach, compared to a female tendency to note ambiguity and seek consensus and commonality between the proffered options (see e.g. Harding, 1979; Murphy, 1982; Ben-Shakhar & Sinai, 1991; Hazel, Logan & Gallagher, 1997; Richardson & O'Shea, 2013).

In recent work, Wilson, Low, Verdon & Verdon (2015) used the questions from eight years of Australian Science Olympiad Examinations in physics to develop a categorisation scheme for MCQs. While individual questions were not often repeated in those examinations, the large number of

students (approximately 600 males and 200 females) attempting each question allowed for the determination of reliable gender gaps. Performance differences in favour of males were largest when questions were 'concrete' (rather than 'abstract') in context, when significant diagrams were present, and when vertical motion, projectiles, or motion in more than one dimension were involved.

Understanding gender gaps allows us to avoid inadvertently creating them, which is of benefit to the educational context generally, as well as in our own institutional context. In this paper, we do not investigate reasons for the overall gender gap in our data, which is likely to be a confluence of many of the factors discussed above. Rather, we identify the characteristics of individual questions that display a gender gap which is (a) significantly larger than the overall/baseline gap for our entire dataset, and (b) persistent over many independent cohorts, from year to year. Questions which have such a large, persistent gender gap are likely to be inappropriate in terms of assessing "ability at physics", as they may well instead be preferentially identifying aspects of gender.

OUR STUDENTS AND THEIR ENVIRONMENT

UNSW Canberra provides tertiary education to officer cadets and midshipmen (together with a relatively small number of commissioned officers, and civilians from the Department of Defence) at the Australian Defence Force Academy. Students are recruited from across Australia, and almost all live on-site while undertaking military training in parallel with their academic studies. The first-year intake is approximately 350 students each year (of which about 25% are female). Attendance at academic activities is compulsory, and students are encouraged to collaborate with each other throughout their academic and military studies. There is significant pressure on students to complete their degree in minimum time so that they can graduate from the Academy together with their peer group. Most students are involved in one or more privately-organised course-specific study groups, where a small number of students collaborate on assignments and tutorial questions.

Each year, approximately 180 students (usually 15%-20% female) take the first-year physics course ZPEM1501 Physics 1A as part of a Science or Engineering degree program. One of the authors (DJL) has taught the first half of Physics 1A (covering kinematics, dynamics, energy and momentum) since 2008, and is responsible for the assessment in that part of the course. Starting from 2010, the assessment format and the degree programs which require the course have all been relatively stable. The course is delivered as three one-hour lectures and one one-hour small-group tutorial each week, together with a fifteen-hour laboratory program. Lectures concentrate on fundamentals, conceptual issues, common misconceptions, and illustrative demonstrations. Tutorials provide a forum for the discussion of how concepts are applied to problems, and highlight problem-solving techniques. The laboratory program aims to teach 'experimental science': while some activities in the laboratory are related to coursework, there is no formal alignment with the lecture/tutorial material, and the laboratory component is assessed separately to the coursework component (via logbook records). Students are given access to a variety of online resources to aid their study, including textbook-specific problem sets: for the period described in this paper, these were *WileyPlus* during 2010-2013, and *Enhanced WebAssign* during 2014-2015.

The first half of the Physics 1A coursework is assessed by two 50-minute 25-question MCQ tests, administered in class, on paper. Each test counts about 17% (± 2%, depending on the year) towards the final mark in the course. Past papers are confidential, and are not available to students. Practice assessment questions in the same style and format as the test questions are circulated, but questions considered to be 'too similar' to those used in formal assessment are avoided. Over the years, a common core of formal assessment questions has been developed, with variation from year to year depending on the timing of the class tests relative to the material, and small variations in emphasis on different aspects of the material. Details of the student cohorts, including performance in each of the two tests, for the years 2010 through 2015 inclusive are presented in Table 1.

OUTLINE OF APPROACH

Since the student cohort and assessment in first-year physics has been reasonably consistent over the years 2010 through 2015 inclusive, we are able to quantitatively evaluate the magnitude and variability of gender gaps over this time, across a large number of MCQs that were each used many times. In this paper, we discuss our findings with respect to those questions that display a statistically consistent gender gap that is significantly greater in magnitude than the overall long-term baseline. We identify the characteristics of questions which display such large, consistent gender gaps. By aggregating student responses across gender and overall performance level, we are able to explore

the ways in which students are answering incorrectly, and suggest what might be causing these gender-based differences in performance. Finally, we make some recommendations for future work to address these issues.

Table 1: details for the two tests (#1 and #2; both 25 MCQs) in ZPEM1501 Physics 1A across 2010-2015. Topic codes indicate the material assessed in each test: K = kinematics, F = forces, D = drag/friction, E = energy/momentum, R = rotation, G = gravitation. Facilities (and gaps) are averaged across the 25 MCQs in each test: standard deviations thus indicate the variability in facility (or gaps) across questions.

Year and Test#	**2010 #1**	**2011 #1**	**2012 #1**	**2013 #1**	**2014 #1**	**2015 #1**
Topics	K,F	K	K,F	K	K,F	K,F
#Male Students	144	159	148	157	156	144
Average Facility	*0.62*	*0.80*	*0.64*	*0.75*	*0.61*	*0.63*
StDev in Facilities	*0.23*	*0.14*	*0.23*	*0.16*	*0.24*	*0.24*
#Female Students	31	31	16	25	30	37
Average Facility	*0.51*	*0.68*	*0.42*	*0.63*	*0.44*	*0.50*
StDev in Facilities	*0.22*	*0.23*	*0.25*	*0.23*	*0.26*	*0.27*
Average Gap in Facility	0.11	0.12	0.22	0.12	0.16	0.12
StDev of Average Gap	0.10	0.11	0.15	0.11	0.11	0.12

Year and Test#	**2010 #2**	**2011 #2**	**2012 #2**	**2013 #2**	**2014 #2**	**2015 #2**
Topics	D,E	F,D,E,R	D,E,R	F,D,E,R	D,E,R	D,E,R,G
#Male Students	147	158	148	158	155	142
Average Facility	*0.55*	*0.58*	*0.63*	*0.57*	*0.58*	*0.61*
StDev in Facilities	*0.19*	*0.19*	*0.19*	*0.21*	*0.18*	*0.22*
#Female Students	35	31	15	23	30	37
Average Facility	*0.47*	*0.51*	*0.45*	*0.46*	*0.46*	*0.52*
StDev in Facilities	*0.19*	*0.19*	*0.20*	*0.24*	*0.21*	*0.24*
Average Gap in Facility	0.08	0.07	0.18	0.11	0.12	0.09
StDev of Average Gap	0.12	0.08	0.14	0.13	0.10	0.08

DATA ANALYSIS

Over the years 2010 through 2015 discussed in this paper, 76 distinct questions have been used in the two Physics 1A tests: 25 questions have appeared in all six years; 6 questions have been used in five of the six years; and 20 questions have been used in four of the six years; the remaining 25 questions have been used in three or fewer of the tests held during that period. Test questions are chosen by the lecturer, to cover core material along with any examples specific to that cohort or year.

For the 51 questions used at least four times (for a total of 260 distinct instances of these questions being asked), the average male facility is 0.65 (standard deviation 0.21), and the average female facility is 0.52 (0.22), with an average gap (defined as [male facility] – [female facility]) of 0.12 (0.08). A histogram of the distribution of gaps for the 260 instances of these 51 questions is shown in Figure 1(a): over 80% of the observed gaps are in favour of males.

To explore the variability in gender gaps, we averaged the facilities for each of the four-to-six instances of each question, for males and females separately, and used the standard deviation in these averages as an indication of the variability. The male and female average facilities for each question are plotted against each other in Figure 1(b). The linear fit to these data is both good (R^2 = 0.86) and nearly parallel to the one-to-one line (slope 0.99), but is offset by the average gap (the *y*-intercept is –0.12).

Our interest in this paper concerns questions which have a consistently large gender gap. Some questions have gaps that vary from instance to instance, which might indicate that individuals within the cohort have a significant effect on the gap from year to year: these questions have large error bars in Figure 1(b). Noting that we have at most six instances of each question, relatively large standard deviations might be expected. Hence, to determine which questions are of particular significance, we have divided each average gap by its standard deviation (σ, calculated from the four-to-six instances of each question). The average gap for 40 questions was less than 2σ; for 7

questions it was between 2σ and 4σ; and for 4 questions it was over 4σ. These last four questions are identified on Figure 1(b), and their significance is discussed in the next section.

As the final component of our data analysis for this paper, we aggregated student answers to each question by gender, both overall and by three overall performance-level groupings. This allows us to examine any differences in answer selection patterns across these factors, and explore variations in the occurrence of particular misconceptions or flawed knowledge structures. The performance-level groupings we used were (i) above or below the average test mark for their own year, (ii) above or below the average test mark for their own gender and year, and (iii) terciles determined by one standard deviation above or below the average test mark for their own year. In the discussion which follows, we draw on all three groupings in order to identify the characteristic answers of sub-cohorts.

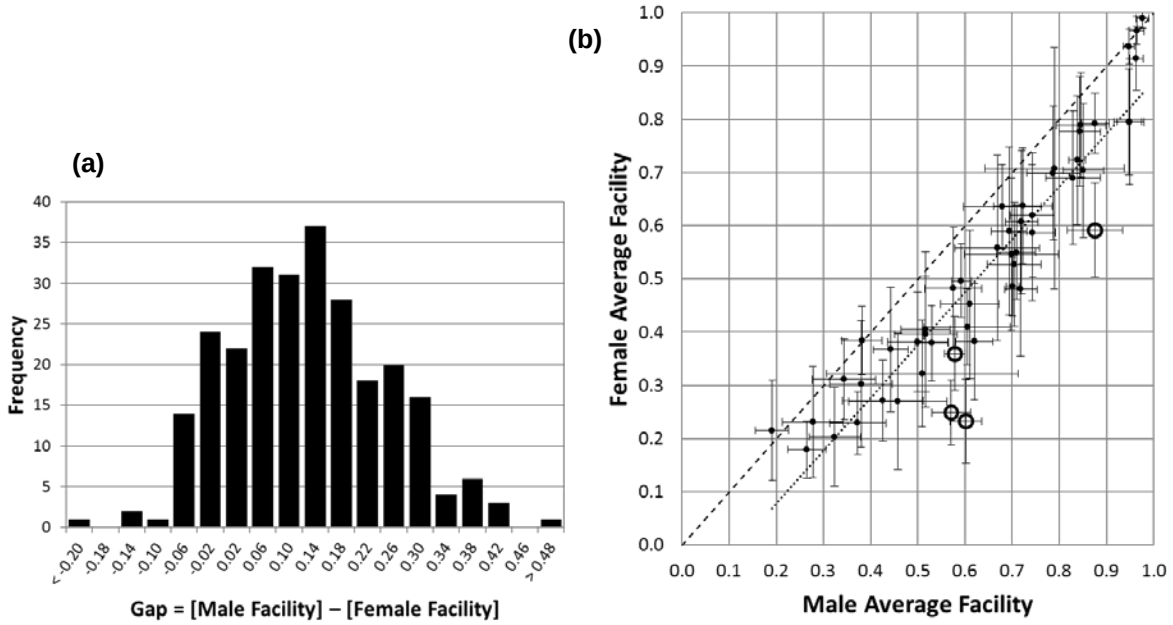

Figure 1: data from the 51 questions which have been used in at least four of the six years of tests between 2010 and 2015. (a) Histogram of gaps from every usage of the 51 questions (260 gaps). Horizontal axis labels give the centre of the bins, which are each 0.04 wide. (b) Scatter plot of average male and female facilities for each of the 51 questions. Error bars are one standard deviation each way. Dotted line is the linear fit to the data points (slope 0.99, *y*-intercept −0.12, and R^2 = 0.86). Points highlighted by open circles are (from left to right): '*Scale*' [0.57, 0.25]; '*Signs*' [0.58, 0.36]; '*Bolt*' [0.60, 0.23]; and '*Raindrops*' [0.88, 0.59].

QUESTIONS OF INTEREST

Of the four questions identified as having large persistent gaps, the question with the largest gap (0.37) between male and female facilities is shown in Figure 2. Under the Wilson et al. (2015) scheme, this question would be expected to have a large gap, as it combines a diagram requiring significant mental gymnastics to interpret with accelerated motion in a real-world context. The gap appears only at the mid- and high-achievement levels, as low-achieving students of either gender struggle with this question. Mid- and high-achieving females display a preference for answers C (indicating difficulty translating the question to the frame of the external observer) and E (indicating difficulty in translating the answer into the frame of the graph's axes).

The question with the most consistent gap is given in Figure 3. Similarly to the previous question, the gap appears predominately at the upper two-thirds of the achievement spectrum, as low-achieving students struggle with it regardless of gender. Overall, females have a stronger preference than males to choose answer A, although the higher-achieving females also exhibit a tendency to choose answer D more often than their male peers. Low achievers of both genders overwhelmingly prefer answer E, although many more low-achieving females than low-achieving males still select answer A. The confluence of 'upwards' and 'downwards' with 'increasing' and 'decreasing' appears to confuse females more than it does males.

An elevator is moving upward with constant acceleration. The dashed curve shows the position y of the ceiling of the elevator as a function of the time t. At the instant indicated by the set of branching (solid) curves, a bolt breaks loose and drops from the ceiling. In the absence of air resistance, which curve best represents the position of the bolt as a function of time **as seen by an external observer**?

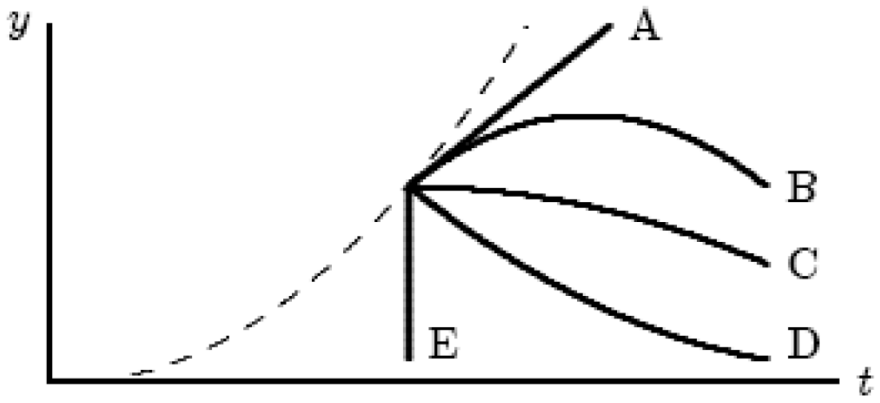

Figure 2: '*Bolt*', used in all six years. Average facilities (standard deviation) are male 0.60 (0.03) and female 0.23 (0.08), with average gap 0.37 (0.09) being 4.2σ.

You stand on a spring scale (e.g. a bathroom or kitchen scale) on the floor of an elevator. Of the following options, the scale shows the **highest** reading when the elevator:

A. moves downward with increasing speed;
B. moves downward with decreasing speed;
C. remains stationary;
D. moves upward with decreasing speed;
E. moves upward at constant speed.

Figure 3: '*Scale*', used in all six years. Average facilities (standard deviation) are male 0.57 (0.04) and female 0.25 (0.06), with average gap 0.32 (0.03) being nearly 12σ.

Throughout a time interval, while the speed of a particle increases as it moves along the x-axis, its velocity and acceleration might be:

A. positive and negative, respectively;
B. negative and positive, respectively;
C. negative and negative, respectively;
D. negative and zero, respectively;
E. positive and zero, respectively.

Figure 4: '*Signs*', used in five years (not 2010). Average facilities (standard deviation) are male 0.58 (0.02) and female 0.36 (0.07), with average gap 0.22 (0.05) being 4.2σ.

Figure 4 shows another of the questions identified in Figure 1(b) as being of particular significance, and it is similar to '*Scale*' (Figure 3) in terms of the conceptual and linguistic mix of 'positive' and 'negative' with 'increasing' speed. While the gap for 'Signs' is not as large as some other questions, it is consistent. The gap here, however, arises from the lower half of the distribution, where it is answered correctly by half of males but only a third of females. Males in the lowest tercile (and females in the middle tercile) prefer option B ('positive acceleration means it's getting faster'), while females in the lowest tercile opt for answers A and E ('positive velocity means speed is increasing').

All three questions discussed above involve significant visualisation: pictures need to be formed in the mind from the given words or diagrams. In addition, words such as 'positive' also need to be uncoupled from their near synonyms to be correctly applied in a physics context to concepts such as directions ('upwards') or changes ('increasing'). Any female advantage in general verbal ability (Halpern et al. 2007; McBride 2009) might actually be a disadvantage in this context as the coupling between near-synonyms is likely to be stronger for females than males. When combined with the visualisation issue, this may explain the large consistent gaps for these questions.

Our final question of interest has much higher average facilities than other questions that display large gender gaps, and is shown in Figure 5. The gap here develops in the lower half of the overall results distribution: over 70% of otherwise low-achieving males answer this question correctly; while low-achieving females tend to answer A or E, with less than 40% choosing the correct answer. We believe that the female advantage in verbal processing discussed earlier is working against these students here, and they may be interpreting the question as 'Why do raindrop**s** fall with **the same** speed…', being cued by the plural into thinking about a large number of drops, and why those drops might be similar (rather than considering why any single drop does not accelerate). The distractors A and C, which imply a comparison between raindrops, further cue students to adopt this misinterpretation.

Why do raindrops fall with near-constant speed during the later stages of their descent?

A. The gravitational force is the same for all raindrops;
B. Air resistance just balances the force of gravity;
C. The drops all fall from the same height;
D. The force of gravity is negligible for objects as small as raindrops;
E. Gravity cannot increase the speed of a falling object to more than 9.8 m s^{-1}.

Figure 5: '*Raindrops*', used in all six years. Average facilities (standard deviation) are male 0.88 (0.06) and female 0.59 (0.09), with average gap 0.28 (0.06) being 4.4σ.

IMPLICATIONS AND FUTURE WORK

Educators should be aware that the wording (particularly undesirable cueing) and presentation of individual questions, as distinct from their content, can lead to large, persistent gender gaps in assessment. Differences in performance between male and female students may be due to gender biases in individual questions, rather than to any overall difference in academic ability. Hence, it is important to review formal assessment material question-by-question for gender gaps, rather than just looking at overall marks. If specific questions are particularly problematic for one gender – for any reason – then (a) we need to be aware of this, and (b) we need to consider what action should be taken to remedy the issue.

In the context of educational research, adding the dimension of gender to any study of student performance is to be encouraged. The work of Meltzer (2005) gives an excellent example of how analysis and reporting of gender differences, while not central to the specific research question, can provide valuable information to other researchers studying parallel topics.

From 2016, we will adjust the wording and presentation of certain questions that have previously displayed consistently large gender gaps, to see if we can improve the facility of female students and thus reduce the gap. For example, we intend asking '*Bolt*' with answer options that are verbal ('The bolt falls straight down', 'The bolt travels up, then down', etc.) rather than pictorial. We will also reword '*Raindrops*' into singular form, in both the question and in the answer options. Post-assessment, we can then test for the statistical significance of any resultant change in gender-based facilities and gap. In doing so, we hope to avoid repeating the findings of McCullough (2004), where a standardised paper revised for gender reduced the gender gap by decreasing the performance of males rather than by improving the performance of females.

ACKNOWLEDGEMENTS

This study was conducted in accordance with the UNSW Canberra Human Research Ethics Advisory Panel approval reference numbers A-13-17 and A-13-37 ("Improving outcomes in first-year physics by the early identification of at-risk students") and A-15-24 ("Identification and closure of gender gaps in physics assessment").

REFERENCES

Bates, S., Donnelly, R., MacPhee, C., Sands, D., Birch, M. & Walet, N.R. (2013). Gender differences in conceptual understanding of Newtonian mechanics: a UK cross-institution comparison. *European Journal of Physics, 34*, 421-434, doi: 10.1088/0143-0807/34/2/421.

Ben-Shakhar, G. & Sinai, Y. (1991). Gender differences in multiple choice tests: the role of differential guessing tendencies. *Journal of Educational Measurement, 28*(1), 23-35, doi: 10.1111/j.1745-3984.1991.tb00341.x.

Dietz, R.D., Pearson, R.H., Semak, M.R. & Willis, C.W. (2012). Gender bias in the Force Concept Inventory? In N.S. Rebello, P.V.Engelhardt & C. Singh (Eds.) *Proceedings of the 2011 Physics Education Research Conference* (pp.171-174). Omaha, Nebraska: American Institute of Physics Conference Proceedings (vol. 1413), doi:10.1063/1.3680022.

Docktor, J. & Heller, K. (2008). Gender difference in both Force Concept Inventory and introductory physics performance. In C. Henderson, M. Sabella & L. Hsu (Eds.) *Proceedings of the 2008 Physics Education Research Conference* (pp.15-18). Melville, New York: American Institute of Physics Conference Proceedings (vol. 1064), doi: 10.1063/1.3021243.

Eurydice (2010). *Gender differences in educational outcomes: study on the measures taken and the current situation in Europe*. Education, Audiovisual and Culture Executive Agency (EACEA P9 Eurydice; Brussels). Retrieved May 26, 2015 from http://eacea.ec.europa.eu/education/eurydice/documents/thematic_reports/120EN.pdf. ISBN 978-92-9201-080-5.

Halpern, D.F., Benbow, C.P., Geary, D.C., Gur, R.C., Shibley Hyde, J. & Gernsbacher, M.A. (2007). The science of sex differences in science and mathematics. *Psychological Science in the Public Interest, 8*(1), 1-51, doi: 10.1111/j.1529-1006.2007.00032.x.

Harding, J. (1979). Sex differences in examination performance at 16+. *Physics Education, 14*, 280-284, doi: 10.1088/0031-9120/14/5/305.

Hazel, E., Logan, P. & Gallagher, P. (1997). Equitable assessment of students in physics: importance of gender and language background. *International Journal of Science Education, 19*(4), 381-392, doi: 10.1080/0950069970190402.

Maccoby, E.E. & Jacklin, C.N. (1974). *The psychology of sex differences*. Stanford, California: Stanford University Press.

Madsen, A., McKagan, S.B. & Sayre, E.C. (2013). Gender gap on concept inventories in physics: what is consistent, what is inconsistent, and what factors influence the gap? *Physical Review Special Topics Physics Education Research, 9* 020121, doi: 10.1103/PhysRevSTPER.9.020121.

McBride, W. (2009). *Teaching to gender differences: boys will be boys and girls will be girls*. Incentive Publications, ISBN 978-0-865307-1-86. Extract available at http://crr.math.arizona.edu/GenderKeynote.pdf.

McCullough, L. (2004). Gender, context, and physics assessment, *Journal of International Women's Studies*, *5*(4), 20-30. Available at http://vc.bridgew.edu/jiws/vol5/iss4/2.

Meltzer, D.E. (2005). Relation between students' problem-solving performance and representational format. *American Journal of Physics, 73* (5), 463-478, doi: 10.1119/1.1862636.

Murphy, R.J.L. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology, 52*, 213-219, doi: 10.1111/j.2044-8279.1982.tb00828.x.

Postles, C. (2013). Girls' learning: investigating the classroom practices that promote girls' learning. In K. Moore, A. Reilly & R. Naylor (Eds.), *Plan UK*, ISBN 978-0-9565219-7-2. Retrieved May 26, 2015 from http://www.plan-uk.org/resources/documents/260260.

Richardson, C.T. & O'Shea, B.W. (2013). Assessing gender differences in response system questions for an introductory physics course. *American Journal of Physics, 81*(3), 231-236, doi: 10.1119/1.4773562.

Wilson, K.F., Low, D.J., Verdon, M. & Verdon, A. (2015). Differences in gender performance on competitive physics selection tests. In review, *Physical Review Special Topics: Physics Education Research*.