REDUCING THE GENDER GAP IN FIRST-YEAR PHYSICS PERFORMANCE

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

Presenting Author: Kate Wilson (k.wilson@adfa.edu.au)
[a]School of Engineering and Information Technology, UNSW Canberra, Canberra, ACT 2610, Australia
[b]School of Physical, Environmental and Mathematical Sciences, UNSW Canberra, Canberra, ACT 2610, Australia

KEYWORDS: gender gaps, physics, first-year

ABSTRACT

Multiple choice tests have been the main means of assessing Newtonian mechanics in the first year physics course at UNSW Canberra for several years. A question-by-question analysis of test results for the years 2010-2015 has previously revealed that some questions had large, persistent gender gaps in performance. In 2016, teaching methods in the course became more interactive and student-centered, and the assessment was modified to include written-answer as well as multiple choice questions (MCQs). Some of the MCQs which had previously shown large and persistent gender gaps were re-used with modifications which aimed to reduce the gap, and others were recast as written answer questions. We observed a small reduction in the overall gender gap. We ascribe this result to better communication: the more interactive teaching methods improved student-student and student-teacher interactions; and the re-casting of assessment questions gave students more means of communicating their understanding to us. However, we note that this improvement comes at the ongoing cost of increased staff time for both facilitating tutorial work and marking.

Proceedings of the Australian Conference on Science and Mathematics Education, The University of Queensland, Sept 28th to 30th, 2016, page 246-253, ISBN Number 978-0-9871834-5-3.

INTRODUCTION

BACKGROUND

Gender gaps in achievement on physics tests have been well documented, and have been observed in many different countries and across age ranges from late primary school to postgraduate study (Madsen, McKagan & Sayre, 2013). These gaps in achievement are of concern because it is very likely that they contribute to the low rate of female participation in Science, Technology, Engineering and Mathematics (STEM) subjects at school and university, and in STEM-based professions. While cultural factors no doubt play a role, if female students underperform on assessment tasks relative to their male peers, this is likely to influence their subsequent choice of study options and hence careers. In spite of efforts at institutional and government level, the gap in participation levels persists and continues to draw high-level attention (Eurydice 2010; Postles 2013).

Halpern, Benbow, Geary, Gur, Shibley Hyde and Gernsbacher (2007) provide an excellent review of the work on gender differences in science and mathematics. The literature on gender gaps in physics is reviewed by Madsen et al. (2013), who conclude 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 (Good, Aronsen & Harder, 2008), self-efficacy (Cavallo, Potter & Rozman, 2004; Sharma & Brewes, 2011), and teacher attitudes (Hazari, Sonnert, Sadler & Shanahan, 2010). Teaching interventions, such as increasing the interactivity of classes, has been found to have a positive effect on the learning of all students (Hake, 1998) and on female students in particular (Lorenzo, Crouch & Mazur, 2006; Postles, 2013), decreasing (but not eliminating) gaps in performance between pre- and post-tests.

When test results are analysed by test item (Wilson, Keuter, Dennis, Nulsen & Verdon, 2007; Morris, Harshman, Branum-Martin, Mazur, Mzoughi & Baker, 2012; Low & Wilson, 2015; Wilson, Low, Verdon & Verdon, 2016), it becomes apparent that there are particular aspects of physics and ways of framing questions that lead to under-performance of female students compared to male students. Projectile motion, whether one or two dimensional, seems to be particularly problematic (Dietz, Pearson, Semak & Willis 2012; Bates, Donnelly, MacPhee, Sands, Birch & Walet 2013; Wilson et al. 2016). Questions in which important information is presented in graphical form have also been found to show gender gaps (Meltzer 2005; Wilson et al. 2016). Multiple-choice questions (MCQs) have been identified by a number of studies as being problematic for females, possibly due to the more 'strategic'

elimination-based approach taken by males, compared to a female tendency to note ambiguity (Hazel, Logan & Gallagher 1997; Richardson & O'Shea 2013).

We have previously described the persistent gaps that we have observed in physics tests (Wilson et al., 2016) and on some particular questions used on our first year physics tests (Low & Wilson, 2015). In the latter work, we identified individual questions that displayed a gender gap which was (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', because they may well instead be preferentially identifying aspects of gender. In this paper, we outline a number of changes that have been made to our first-year physics course structure, to the assessment mix within that course, and to the nature of individual assessment questions. We discuss the results of these modifications, and consider what insight they may give us into means by which gender bias in physics testing may be reduced.

OUR STUDENTS AND THE PHYSICS COURSE

UNSW Canberra provides tertiary education to officer cadets and midshipmen (together with a small number of commissioned officers and civilians) 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. Approximately 180 of these 350 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. From 2010 to 2015 the assessment format and the degree programs which require the course were relatively stable. More details of the course structure and learning activities during 2010-2015 can be found in Low and Wilson (2015). During those years, the first half of the Physics 1A coursework was assessed by two 50-minute 25-question MCQ tests, administered in class, on paper. Each test counted about 17% (± 2%, depending on the year) towards the final mark in the course. Past papers are confidential, and are not available to students.

In 2016, there were substantial modifications to the delivery of the course. Instead of three hours of lectures and one hour of tutorial, the students were taught through two hours of lectures and one two-hour workshop-tutorial each week. Instead of one academic tutor being responsible for multiple tutorial classes each of about 16 students, three academic tutors worked together in a room with 50-60 students. Within those larger tutorial classes, participation became more interactive, with small groups (4-6 students) working around a whiteboard. Simple hands-on activities (see e.g. Wilson, Sharma, Millar, Moroney, Newbury, Logan, Cathers, Vella & Emeleus, 2002) were included in the tutorials to illustrate concepts and supplement theoretical problem-solving. In comparision to the previous tutorial structure, the students spent more time actively exploring and communicating their understanding of the physics both with each other and with the teaching staff. This was done in response to institutional pressure to move away from lectures as the main means of teaching, to a partially "flipped mode" course structure, as part of UNSW teaching and learning policy. The fraction of MCQs in the assessment was reduced, with a single class test consisting of 12 MCQs and 4 written (problem solving) questions replacing two 25-MCQ papers.

The 2016 cohort was of similar size and composition to previous years. As UNSW Canberra draws students from across Australia, there is no reason to expect that curriculum changes (if any) in a given state will have significant effect on the performance of the cohort overall. We have noted that there are no significant differences in performance by students from different states. Further, the 2016 cohort's pre-course performance on the Force Concept Inventory (Halloun, Hake, Mosca and Hestenes, 1995) was not different to previous years' cohorts.

OBSERVATIONS

Four of the 12 MCQs used in the 2016 class test were previously identified by Low and Wilson (2015) as showing large, consistent gender gaps over the years 2010-2015. Two of these questions (*Bolt*, Figure 1; and *Raindrops*, Figure 3) were revised for 2016, in an attempt to reduce the gender gap, while the other two were re-used unchanged. Three other MCQs from the 2010-2015 physics class tests were recast into written answer problems in 2016. All three of these problems required some

calculation, and had a numerical answer: *Ferris-wheel* involved a combination of uniform circular motion and projectile motion; *Sled* involved the setup and analysis of a free-body diagram using Newton's Second Law; and *Spring* involved determining the work done in extending a spring.

To analyse the differences between 2016 and previous years, we will consider each component of the 2016 test separately. The numerical data are presented in Table 1.

UNCHANGED MULTIPLE CHOICE QUESTIONS

The ten MCQs which were re-used unchanged in 2016 give us an indication of how effective the changed mode of teaching is for the content examined by these questions. The aggregated change in facility (the fraction of a particular cohort that answered correctly) on these questions does not provide any evidence of improvement by the male students. However, there is a borderline significant improvement in female performance, resulting in a significant reduction in the gender gap. Hence it seems likely that this mode of instruction, in which there is more interpersonal communication among students in class time, as well as more one to one communication between teaching staff and students, has a beneficial effect on females in particular. This is consistent with previous findings, that interactive teaching methods improve the performance of female students, and thus reduce the gender gap (Lorenzo et al. 2006; Brew & Ginns 2008).

Table 1: facilities and gaps for the two modified MCQs (*Bolt* and *Raindrops*), the three written answer questions previously given as MCQs (*Ferris-wheel*, *Sled* and *Spring*) and the 10 MCQs used unchanged in 2016. In 2016, the cohort consisted of 136 males, 30 females. For 2010-2015, cohort details are in Table 1 of Low & Wilson (2015); average of 151 males and 28 females. Differences of more than one standard deviation (sd) are significant at the 95% level.

Question	2010−2015 facility *(sd/sem)*			2016 facility			Change in facility/gap		
	Male	Female	Gap	Male	Female	Gap	Male	Female	Gap
Bolt	0.60 *(0.03)*	0.23 *(0.08)*	0.37 *(0.09)*	0.43	0.33	0.10	−0.17	+0.10	−0.27
Raindrops	0.88 *(0.06)*	0.59 *(0.09)*	0.28 (*0.06)*	0.76	0.50	0.26	−0.12	−0.09	−0.02
Ferris-wheel	0.50 *(0.06)*	0.38 *(0.09)*	0.12 *(0.12)*	0.60	0.55	0.05	+0.10	+0.17	−0.07
Sled	0.32 *(0.05)*	0.20 *(0.09)*	0.12 *(0.10)*	0.37	0.30	0.07	+0.04	+0.10	−0.05
Spring	0.38 *(0.07)*	0.30 *(0.12)*	0.08 *(0.11)*	0.39	0.39	0.00	+0.01	+0.08	−0.08
Unchanged MCQs (10)	0.56 *(0.08)*	0.43 *(0.08)*	0.13 *(0.03)*	0.58 *(0.07)*	0.50 *(0.08)*	0.08 *(0.03)*	+0.02	+0.07	−0.05

MULTIPLE CHOICE QUESTIONS CONVERTED TO WRITTEN ANSWER QUESTIONS

Three questions (*Ferris-wheel*, *Sled* and *Spring* in Table 1) were converted from the MCQ format employed over 2010-2015, to a written-answer format in 2016. As written questions, each was marked out of 8: 2 marks for *S*etting up the problem, including a diagram; 2 marks for identifying the relevant *P*hysics approach and noting any assumptions made in applying this approach; 2 marks for setting up and rearranging the *E*quations for the quantity of interest; and 2 marks for obtaining the correct *N*umerical answer, and expressing it with the correct units and an appropriate number of significant figures. Dividing the total by 8 thus converts these marks to an equivalent facility.

As practice for the written-answer section of the Class Test, students sat weekly tutorial quiz problems (the best ten of which counted 19% towards the final grade in the course). These quizzes were presented in the same format, and marked using the same criteria as the Class Test written-answer questions. Students were reminded, both before and after each weekly quiz how these questions would be marked, and were given the marking criteria explicitly as a rubric on each question paper. However, it was noted in in formal and informal discussions with students that many students still expected to be given full marks for a correct numerical answer, and were prepared to argue that the marking had been unfair if a lack of set-up or working resulted in a lower mark.

As MCQs, these three questions each had a moderate gender gap, with an average gap of 0.11. As written questions, we find both male and female performance is improved over the MCQ version, and the average gap decreases to 0.04 (due to the female facility increasing more than the male facility).

A closer examination of the marks breakdown (shown in Table 2) revealed that female students (on average) scored higher for Setting up the problem and drawing diagrams, but lower on each of the other three criteria. The reasons for these differences are currently a matter of conjecture, and require further study. For example, our male students may be putting less effort into drawing useful diagrams because they are less diligent, because they see do not see the value in drawing diagrams of situations they are able to visualise, or for some other reason. Hence letting students communicate their understanding to us in different ways has indicated gender differences here too.

Table 2: average mark (out of 2) by gender within each assessment category for the four written answer questions used in 2016.

	Setting Up		Physics Approach		Equations		Numerical Solution	
	Male	Female	Male	Female	Male	Female	Male	Female
Ferris-wheel	1.00	1.03	1.35	1.17	1.16	1.08	1.31	1.12
Sled	0.87	0.92	0.68	0.55	0.76	0.52	0.64	0.43
Spring	0.93	1.08	0.75	0.80	0.88	0.68	0.61	0.53
Railcar	0.46	0.63	0.54	0.39	0.49	0.26	0.13	0.05

MODIFIED MULTIPLE CHOICE QUESTIONS

The question identified by Low and Wilson (2015) as having the largest gap for the 2010-2015 cohorts is shown in Figure 1. In order to test the hypothesis that the gender gap arose due to the way in which information was being communicated to students (as distinct from it being somehow inherent to the content), this question was changed for 2016 to the version shown in Figure 2, where the graphical presentation has been replaced by a textual description. The distractors were written to be as close as possible to the graphical versions in Figure 1, although no direct translation of option E was considered plausible. With the removal of the need to interpret the graph, the facility for female students increased from 0.23 to 0.33. At the same time, however, the facility for male students decreased from 0.60 to 0.43. The overall result is a narrowing of the gap from 0.37 to 0.10, although not entirely for the reasons one might hope. Similar results were observed by McCullough (2004) when attempting to decrease gender gaps by rewording questions.

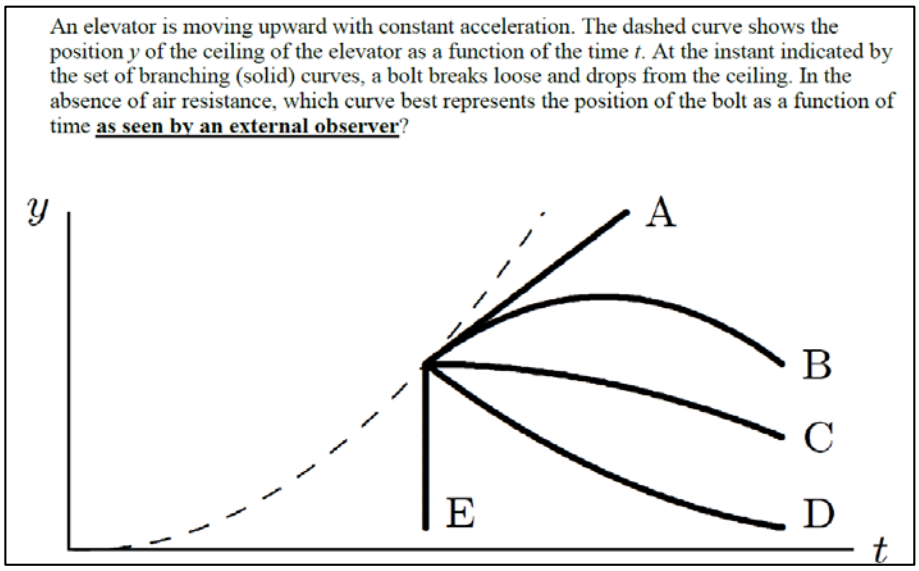

Figure 1: '*Bolt*', used in 2010-2015 inclusive. Average facilities (standard deviation) were male 0.60 (0.03) and female 0.23 (0.08), with average gap 0.37 (0.09).

More information can be gained by looking at differences in the distribution of student answers to these two versions of the question, which are shown in Table 3. In the absence of the graphical clues, and relying on words alone, the males in 2016 were less likely to correctly identify the 'up and down' trajectory. Removing the difficulty with translating the motion into a graphical representation appears to have increased female facility in 2016. However, option C of the worded options draws out a difficulty in translating between reference frames, particularly for female students.

A glass-walled elevator is moving upwards, with constant acceleration. At some point in the elevator's motion, a bolt breaks loose and drops from the ceiling. What is the motion of the bolt as seen by an **external** observer (i.e. one located outside the elevator)?

A. The bolt moves upwards at constant speed;
B. The bolt first moves upwards, then reverses direction and moves downwards;
C. The bolt appears to remain stationary;
D. The bolt immediately moves downwards, accelerating under gravity;
E. The bolt immediately moves downwards, at constant speed.

Figure 2: '*Bolt-2*', used in 2016. Facilities were male 0.43 and female 0.33 (gap 0.10).

Table 3: the percentage of each cohort that picked each answer-option for the multiple choice questions *Bolt* (2010-2015) and *Bolt*-2 (2016), by gender. Options A, B and E are equivalent between the two tests, while *Bolt* C is equivalent to *Bolt*-2 D. While *Bolt* D and *Bolt*-2 C are not equivalent, for convenience they are displayed in the same column.

	Bolt (graphical, 2010-2015) / Bolt-2 (textual, 2016)				
Gender	**A / A**	**B / B**	**C / D**	**D / C (n.e.)**	**E / E**
Male	4 / 1	60 / 43	19 / 32	9 / 22	8 / 1
Female	8 / 0	24 / 33	27 / 17	21 / 40	20 / 10

The second question that was modified has much higher average facilities than other questions that display large gender gaps. The original version is shown in Figure 3, where in 2010-2015, the 12% of male and 41% of female students who answered incorrectly were approximately evenly split between distractors A and E. Based on this, Low and Wilson (2015) hypothesised that females in particular may have been interpreting the question as 'Why do raindrop**s** fall with **the same** speed…', being cued by the plural in the question's wording 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. Distractor C, which implies a comparison between raindrops, may 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 3: '*Raindrops*', used in 2010-2015 inclusive. Average facilities (standard deviation) were male 0.88 (0.06) and female 0.59 (0.09), with average gap 0.28 (0.06).

Why does a raindrop fall with near-constant speed during the later stages of its descent?

A. The gravitational force is constant;
B. Air resistance just balances the force of gravity;
C. The height from which the raindrop started falling is fixed in space;
D. The force of gravity is negligible for objects as small as a raindrop;
E. Gravity cannot increase the speed of a falling object to more than 9.8 m s^{-1}.

Figure 4: '*Raindrops-2*', used in 2016. Facilities were male 0.76 and female 0.50, with gap 0.26.

In an attempt to remove this possible source of confusion, the question was reworded for the 2016 class test as shown in Figure 4. Unfortunately, this had no impact on the size of the gap at all, and if anything reduced both male and female facilities compared to the original version. Males still chose A and E as their distractors of preference in the revised version, although at a 2:1 ratio rather than an even split. Most (11/15) of the females in 2016 who answered incorrectly chose A. The answer A is a true statement in the 2016 version (Figure 4), but not in the older version (Figure 3). However, answer A is not the answer to the question being asked, so it may be that female students in particular are choosing that distractor for different reasons now. If females are looking for something to justify an answer, rather than to eliminate it, then this becomes an appealing option.

In preliminary interviews, for example, one female student chose A on the grounds that it is a true statement, and rejected the correct answer (B) as 'it doesn't sound scientific'. Subsequent interviews and discussions showed that the wording of the correct answer may be problematic for many female students: in particular, the word 'just' discouraged them from selecting the correct answer. In contrast, male students generally did not even notice the inclusion of the word 'just' in option B. While the content of the question may well contribute to the gap (many researchers, including Docktor and Heller (2008), Dietz et al. (2012) and Bates et al. (2013), have reported large gender gaps on questions involving projectiles and objects falling due to gravity), at least part of the gap appears to be due to the written communication of the question. Clearly, further investigation is warranted.

IMPLICATIONS AND FUTURE WORK

Changing the way in which we teach to encourage communication among students, and between students and teaching staff, provides opportunities to share different 'ways of knowing'. While this approach appears to be effective for improving learning, particularly for females, it comes at the cost of ongoing increased staff time, as well as the large initial time investment in modifying the course.

The gender gap can be narrowed by changing the format of assessment questions from multiple-choice to written answer. This has additional benefits to all students in that allows them to demonstrate their depth of knowledge, and may change study habits to encourage deeper learning (Mullen and Schultz, 2012) However, the cost of such a change has been a large increase in staff time spent marking. Where classes are large, or staff time limited, this may not be a viable option.

Changing the way in which a question is presented can also narrow (or widen) the gender gap. The original *Bolt* question (Figure 1) showed a large gap, due to a combination of the question's content (projectile motion/object falling under gravity) and the way in which that content was communicated (graphs). When the method of communication was changed, the gap decreased. The gap due to the nature of the content remained.

We are not suggesting that questions requiring students to interpret graphs should not be asked: this is a basic skill in science. Nor are we suggesting that projectile motion not be examined. When the two are combined, however, this is likely to place additional cognitive load on females in particular, given the widespread underperformance of female students on questions involving the graphical presentation of information, and on questions involving projectile motion (Wilson et al. 2016). It also means that, unless the distractors are carefully written and the choice of distractor interrogated, it will not be possible to say whether a student who answered incorrectly did so because they couldn't read the graph, or because they lacked understanding of projectiles. Hence, the question is less useful as a diagnostic test of student understanding - one of the main purposes of any assessment task.

Further, we would argue that requiring students to interpret words is also a necessary skill. Scientists often have to deal with non-scientists, who generally communicate in words rather than graphs. Hence, being able to interpret words is as necessary a skill as being able to interpret graphs, and arguably a more transferable one. We also note that it is not always easy to predict what effect changing the way the question is communicated will have. The re-worded *Raindrops* question (Figure 4) showed just as large as gap as the original version (Figure 3). While it is not feasible to require academics in general to carry out psychometric analyses of their assessment tasks, as for example Barbera (2013) has done for the chemistry concept inventory, careful consideration of precisely what is being tested, and whether the test items best meet this need, should be considered.

To conclude, we suggest that educators should be aware that communication, including the way information is communicated on tests, can contribute to gender gaps in performance. We need to consider not only how we communicate with students in the way that we present information, but also how we ask them to communicate their understanding to us: by choosing one of a set of options that we have provided, or by drawing their own pictures and writing their own words. If specific questions or types of questions are particularly problematic for one gender for any reason, then we need to be aware of this. We can then choose how we respond: by adjusting our assessment practices, and/or by modifying our teaching practices to provide support as needed.

ACKNOWLEDGEMENTS

This study was conducted in accordance with the UNSW Canberra Human Research Ethics Advisory Panel approval reference numbers A-13-17, A-13-37 and A-15-24.

REFERENCES

Barbera, J. (2013). A psychometric analysis of the chemical concepts inventory. *Journal of Chemical Education, 90*, 546-553, doi: 10.1021/ed3004353.

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.

Brew, A. & Ginns, P. (2008). The relationship between engagement in the scholarship of teaching and learning and students' course experiences. *Assessment & Evaluation in Higher Education, 33*, 535-545, doi: 10.1080/02602930701698959.

Cavallo, A. M., Potter, W. H., & Rozman, M. (2004). Gender differences in learning constructs, shifts in learning constructs, and their relationship to course achievement in a structured inquiry, year long college physics course for life science majors. *School Science and Mathematics*, *104*(6), 288-300.

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.

Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology*, *29*(1), 17-28.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64-74.

Halloun, I., Hake, R., Mosca, E., & Hestenes, D. (1995). Force Concept Inventory (revised 1995). Retrieved from http://modeling.asu.edu/R&E/Research.html (password protected).

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.

Hazari, Z., Sonnert, G., Sadler, P. M., & Shanahan, M. C. (2010). Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: a gender study. *Journal of Research in Science Teaching*, *47*(8), 978-1003.

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.

Lorenzo, M., Crouch, C. H., & Mazur, E. (2006). Reducing the gender gap in the physics classroom. *American Journal of Physics*, *74*(2), 118-122.

Low, D. J. & Wilson, K. F. (2015). Persistent gender gaps in first-year physics assessment questions, In M. Sharma & A. Yeung (Eds.) *Proceedings of the 2015 Australian Conference on Science and Mathematics Education* (pp. 118-124). Perth, Australia, 30th September – 2nd October 2015, http://openjournals.library.usyd.edu.au/index.php/IISME/article/view/8775.

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.

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.

Morris, G. A., Harshman, N., Branum-Martin, L., Mazur, E., Mzoughi, T., & Baker, S. D. (2012). An item response curves analysis of the Force Concept Inventory. *American Journal of Physics*, *80*(9), 825-831, doi: 10.1119/1.4731618.

Mullen, K.; Schultz, M. Short answer versus multiple choice examination questions for first year chemistry (2012). *International Journal of Innovation in Science and Mathematics Education*, *20*, 1-18.

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.

Sharma, M. D. & Bewes, J. (2011). Self-monitoring: Confidence, academic achievement and gender differences in physics. *Journal of Learning Design*, *4*, 1-13, doi: 10.5204/jld.v4i3.76.

Wilson, K., Sharma, M., Millar, R., Moroney, C., Newbury, R., Logan, P., Cathers, I., Vella, G., & Emeleus, G. (2002). *Workshop tutorials for physics*. The University of Sydney, NSW: UniServe Science.

Wilson, K., Kueter, N., Dennis, G., Nulsen, A., & Verdon, M. (2007). Addressing gender disparity in the Physics National Qualifying Exam for the Australian Science Olympiads. *Teaching Science*, *53*(1), 24-29, ISSN 1449-6313.

Wilson, K. F., Low, D. J., Verdon, M., & Verdon, A. (2016). Differences in gender performance on competitive physics selection tests. *Physical Review Physics Education Research, 12*(2) 020111, doi: 10.1103/PhysRevPhysEducRes.12.020111.