DIN YAN YIP, MING MING CHIU and ESTHER SUI CHU HO

HONG KONG STUDENT ACHIEVEMENT IN OECD-PISA STUDY: GENDER DIFFERENCES IN SCIENCE CONTENT, LITERACY SKILLS, AND TEST ITEM FORMATS

ABSTRACT. This study examined gender differences in students' scientific literacy as measured by OECD-PISA. In particular, we focused on the 2437 students from 140 Hong Kong schools. Hong Kong boys' and girls' science scores did not differ overall. However, boys scored higher than girls at the higher percentiles (75th and above). Moreover, specific test components showed gender differences. Boys tended to score higher on tests with more earth and physical science items, understanding of scientific knowledge items, and closed items. Meanwhile, girls tended to score higher on 'recognizing questions' and 'identifying evidence' items. These results suggest that a science test assessing diverse content and literacy skills in a variety of response formats provides both a more comprehensive picture of students' capabilities and a more likely gender-equitable assessment.

THE ISSUE OF GENDER DIFFERENCES IN SCIENCE EDUCATION

Since the 1970s, a number of international surveys have provided evidence of gender differences in science achievement across different nations or regions. For example, two IEA studies (International Association for the Evaluation of Educational Achievement) in the 1970s and 1980s reported consistent gender differences in science performance that favoured the boys, and the differences increased with age of schooling (Keeves, 1986, 1992). A similar pattern of performance was also reported in a survey conducted by the USA National Assessment and Educational Progress (NAEP, 1978). In a more recent IEA study, the Third International Mathematics and Science Study (TIMSS), similar gender differences in science performance were identified in many of the participating countries, and the performance gap became more pronounced at higher grade levels (Law, 1996; Mullis, Martin, Beaton, Gonzalez, Kelly & Smith, 1998). The differential performance between boys and girls has been linked to the dominance of boys over girls in entering science courses in the higher secondary and university levels, particularly in physical sciences (Gorard, Salisbury & Rees, 1999; Head, 1999).

A variety of reasons have been offered to account for such gender differences in science performance, and they raise concerns about the issue of

International Journal of Science and Mathematics Education **2:** 91–106, 2004.

equity in science education. Browne and Ross (1991) and Murphy (1997) noted that boys and girls are different in their interests and expectations from an early age. These gender differences may shape the children's perceptions of self-competence in various school subjects, which may in turn affect their achievements in science. Murphy (1991) found that girls tend to consider contextual features as an integral part of the science tasks while boys tend to consider issues in isolation. Thus girls usually formulate more complex multivariable investigations that are difficult to work on, but the difficulty is often interpreted by teachers as evidence of girls' misunderstanding or incompetence in science.

Boys and girls also differ in their styles of learning. Gorman, White, Brook, Maclure & Kispal (1988) showed that at age 15, more boys than girls prefer reading books that give accurate facts, while more girls like to read to help understand their own and other people's personal problems. Kimbell, Stables, Wheeler, Wosniak & Kelly (1991) showed that girls prefer working in collaboration through discussion with others, while boys prefer working independently and quickly. Thus boys' learning style will be favoured by a more traditional style of science teaching featured by lecturing and teacher explanation with relatively little classroom interaction (Murphy, 1999), which is particularly prevalent in science lessons at higher secondary levels.

In recent years, research findings have questioned the gender equity of traditional assessment practices, which may favour the forms of knowledge and ways of knowing that are more likely to be acquired by boys than by girls. There is evidence that boys tend to perform better than girls on timed, competitive, external tests and girls work better on cumulative, non-competitive, school-based assessment (e.g., Blithe, Clark & Forbes, 1994; Hildebrand & Allard, 1993; Parker & Tims, 1994). Furthermore, boys tend to perform better in topics related to earth science and physical science, and girls in topics involving health and nutrition (Mullis, Martin, Fierros, Goldberg & Stemler, 2000). Research findings also suggest that boys are favoured by multiple-choice tests and girls by extended-response items (e.g., Harding, 1979; Hoste, 1982). However, the study of Jovanovic, Solano-Flores & Shavelson (1994) indicated an advantage to boys on multiple-choice tests for a physical science topic but not for a biological science topic. Analyses of large data in the USA (Kahle & Lakes, 1983) and UK (Johnson, 1987; Murphy, 1991) revealed that there were gender differences in science performance according to whether the test items were set in contexts typical of male or female backgrounds and experiences.

The above findings show that using a narrow range of assessment tasks and strategies may yield a gender-biased picture of students' science capabilities. Comprehensive and equitable assessment entails diversity in both the content and form of assessment. These include the use of school-based and external tests, which can be competitive or non-competitive in nature, and a balanced distribution of items on both physical and biological sciences. Test items should assess scientific knowledge as well as other cognitive skills, and should be set in different contexts and formats that demand both close-ended and open-ended responses.

This paper examines the gender differences in science performance of Hong Kong students in a large-scale international survey conducted by the Organisation for Economic Co-operation and Development (OECD). This project, known as the Programme for International Student Assessment (PISA), aims at assessing students' achievement in scientific literacy across different countries and regions. The PISA survey is therefore different from most other international studies that focus mainly on assessing students' mastery of scientific knowledge in common curriculum areas of different countries.

The following sections will start with a brief introduction of the rationale and design of the PISA project, with particular reference to the meaning of the scientific literacy framework and the nature of test items used for assessing scientific literacy. Next, we analyse the performance of the girls and boys of the Hong Kong sample, including both the overall performance and performance in different components of the scientific literacy framework. The final section discusses the assessment strategy used in the scientific literacy framework of PISA and explores the potential of this strategy for making equitable assessments of the achievement of girls and boys in science learning.

THE PISA PROJECT

Aims of the PISA Study

OECD's PISA project assesses students' competence at using their knowledge and skills to solve problems and make informed decisions in everyday life situations. The term "literacy" is used to describe the ability of this nature, to distinguish it from the ability to recall and understand subject matter knowledge from the school curriculum. The first survey of 15-year-olds' literacy in reading, mathematics and science (PISA 2000) was conducted in 32 countries (including 28 OECD countries) in 2000.

Nine additional countries or regions, including Hong Kong, conducted the survey in 2002.

The international results of PISA 2000 are published in two reports (OECD, 2001, 2003), and the results for the Hong Kong sample are elaborated in a separate report (HKPISA, 2003). Building on the analyses presented in these reports, this paper explores a particular aspect of scientific literacy of Hong Kong students: the gender differences in science performance.

The Scientific Literacy Framework

Scientific literacy by age 15 is a key education goal, whether or not students continue their study of science. In the PISA project, scientific literacy is defined as "the capacity to use scientific knowledge, to identify questions and to draw evidence-based conclusions in order to understand and help make decisions about the natural world and the changes made to it through human activity." (OECD, 2001, p. 23.)

Scientific literacy is more than recalling scientific facts and vocabulary. Scientific literacy encompasses understanding of scientific knowledge, the processes for developing this knowledge, and the nature of scientific knowledge. Accordingly, PISA 2000 designed tasks to assess the following components:

1. Ability to demonstrate understanding of scientific concepts.
2. Ability to recognise scientifically investigable questions.
3. Ability to identify evidence needed in a scientific investigation.
4. Ability to draw or evaluate conclusions.
5. Ability to communicate explanations or conclusions.

PISA 2000 assesses the thematic areas of *Earth and environment*, *Life and health* and *Science in technology*, rather than traditional subjects such as physics, chemistry and biology. Unlike the traditional subject areas, these themes are more relevant to people's everyday life and more in line with PISA's view of scientific literacy as a prerequisite for adult life.

The assessment tasks set within these themes are all extended tasks, not isolated single item tasks. They include items assessing understanding of scientific concepts (44%), recognizing questions suitable to scientific investigation (15%), identifying evidence (15%), drawing or evaluating conclusions (18%) and communicating conclusions (9%) (Table I). There were also closed (65%) and open (35%) items. Closed items include multiple-choice, true-false, matching items, and short answers (answered with a few words). Closed items assess understanding of basic scientific knowledge, with little demand on conceptual integration or communica-

TABLE I

Distribution of assessment items for the scientific literacy framework

	Item types and number of items		
	Closed items	*Open items*	*Total*
Distribution of items by abilities			
To demonstrate understand of scientific concepts	12 (13)	3 (3)	15 (16)
To recognise scientifically investigable questions	4 (4)	1 (1)	5 (5)
To identify evidence needed in a scientific investigation	3 (3)	2 (2)	5 (5)
To draw or evaluate conclusions	3 (3)	3 (4)	6 (7)
To communicate valid conclusions		3 (5)	3 (5)
Distribution of items by thematic areas			
Earth and environment	7 (8)	6 (8)	13 (16)
Life and health	8 (8)	5 (6)	13 (14)
Technology	7 (7)	1 (1)	8 (8)
Total	22 (23)	12 (15)	34 (38)

*The number inside the brackets indicates the scores allocated to the items.

tion skills. Open items require extended responses and demand higher order skills such as evaluation and integration as well as communication skills.

Design of the Study

Sampling Procedure and Data Collection

The students were selected by a two-stage stratified sampling design. In the first stage, schools were sampled systematically with probabilities proportional to the number of 15-year-old students enrolled. In Hong Kong, 150 schools were selected, and 140 schools were included in the final data set. In the second stage, 35 students were randomly selected from the list of 15-year old students in each sampled school. Of these students, 4405 completed the tests. However, the anchor-test design (Lord, 1980) included science questions in only 5 of the nine booklets, so only 2437 students completed the science portions of the tests. Each student worked on an assessment booklet for one and a half hours in his/her own school. Details of the sampling procedure and assessment design for Hong Kong students are described in the Hong Kong regional report (HKPISA, 2003).

Analysis

OECD (2002) analysed the test scores by fitting a graded response Rasch model to the data. Graded response Rasch allows for missing data, models test item difficulty and partial credit answers. As each participant was only given a portion of the entire test to complete, a Rasch model computes their achievement scores based only on the questions they receive. As the difficulty of each item differs, a Rasch model estimates the difficulty of each item to yield more precise student achievement scores (Lord, 1980). Furthermore, the graded response aspect of the model captures the partial credit given to some answers (Samejima, 1969). OECD (2002) fitted this Rasch model to create weighted maximum likelihood estimates (Warm, 1985). OECD (2002) then rescaled these achievement scores to have an overall mean of 500 and a standard deviation of 100 across participants in all OECD countries. So, about 68% of the participants from OECD countries scored between 400 and 600 points.

RESULTS AND DISCUSSION

Gender Differences in Scientific Literacy in PISA 2000 Countries

Of the 41 countries and regions tested, boys performed significantly better than girls in three countries, i.e., Korea, Denmark and Austria, whereas girls outperformed boys in six countries, Albania, Latvia, Macedonia, New Zealand, Thailand and Russia (Figure 1). Boys and girls did not significantly differ in the other 32 regions, including Hong Kong.

According to these results, there appears to be no consistent pattern of gender difference in scientific literacy among different types of countries, such as between the ‘high performance’ and ‘low performance’ countries, between the OECD and non-OECD countries, or between countries with Asian and Western cultures. So, gender differences are likely to be due to country-specific factors, such as curriculum, learning environment or societal context.

These results differ from those of other international studies. For example, the TIMSS survey of 8th grade students aged 14 reported that in most participating countries, boys scored higher than girls in science (Law, 1996). In the TIMSS study, Hong Kong showed the largest gender difference in science performance in favour of boys at the 8th grade among all participating countries that satisfied the sampling and participation requirements for international comparison. Such a gender difference is, however, not shown by Hong Kong students in the present study.

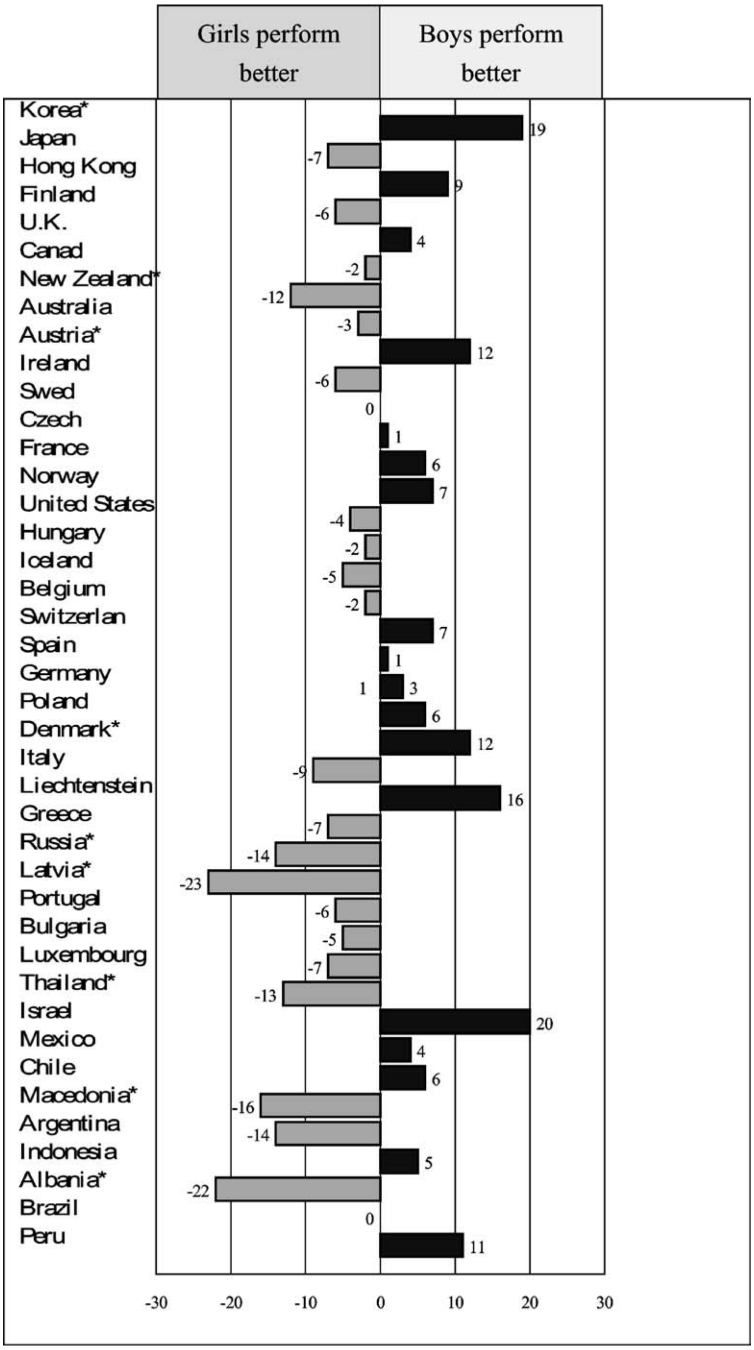

Figure 1. Gender differences in performance on scientific literacy in PISA 200 countries, sorted by country's mean performance from highest (Korea) to lowest (Peru).
*Countries in which gender difference in performance is statistically significant.

A possible reason for the discrepancy in gender effect on science achievement as identified by PISA and TIMSS is that the two international studies are assessing different aspects of science achievement (OECD, 2001, p. 126). The PISA assessment of scientific literacy places greater emphasis than TIMSS on life science, in which girls tend to perform better than the boys. TIMSS, on the other hand, focuses on students' scientific knowledge as prescribed in the national science curriculum that usually has a greater emphasis in physical science, in which boys tend to perform better than the girls (Schmidt, Raizen, Britton, Bianchi & Wolfe, 1997). Comparing with TIMSS, PISA has a higher proportion of open-response and contextualised items, in which girls tend to perform better, rather than multiple-choice items which may favour the boys (Jovanovic et al., 1994; Volkoff & Hocevar, 1995; Whitehouse & Sullivan, 1992). According to the scientific literacy framework, test items in the PISA study are designed to assess a general understanding of the important concepts of science, the methods of science, the nature of scientific knowledge, and the strengths and limitations of science in everyday life. These abilities are believed to be essential for future life in society. The test items in TIMSS, on the other hand, are concerned with the mastery of scientific knowledge and skills that are essential for pursuing further studies in science. These differences in emphasis of assessment between PISA and TIMSS may account for the different observations on gender effect in science achievement for Hong Kong students obtained from these two international studies. The validity of some of the suggested reasons will be examined on the basis of the data collected in the present study in subsequent sections.

Gender Differences in Scientific Literacy Across Different Ability Levels

Hong Kong students' science scores were among the best in the world with a mean (M) score of 541 and a standard error (SE) of 3. No other country scored significantly higher. So, the analyses below examine gender differences in a high science achievement region.

Overall, Hong Kong boys did not significantly outperform Hong Kong girls in scientific literacy (Figure 1). In the higher percentiles (75th and above) however, boys scored higher than girls (Table II). This result is consistent with the greater percentage of boys in the science stream of senior secondary years in Hong Kong. The better performance of the boys in the higher ability groups of the 15-year-olds suggests that they can compete more successfully with the girls for the limited number of science places in the S4 level (equivalent to grade 10), the first year for assigning students into science and non-science streams. This in turn will lead to the

TABLE II

Hong Kong girls' and boys' scores in scientific literacy at different percentiles

	Female		Male		Difference of
	Mean	S.E.	Mean	S.E.	the means
95th percentile	659	5.9	680	7.1	−21*
90th percentile	635	4.4	655	6.2	−20*
75th percentile	592	4.4	608	5.3	−15*
50th percentile	536	3.7	545	5.0	−9
25th percentile	486	5.2	489	6.9	−3
10th percentile	429	9.5	423	8.1	6
5th percentile	393	9.4	386	10.2	8

*Difference of the means is significant at the 0.05 level.

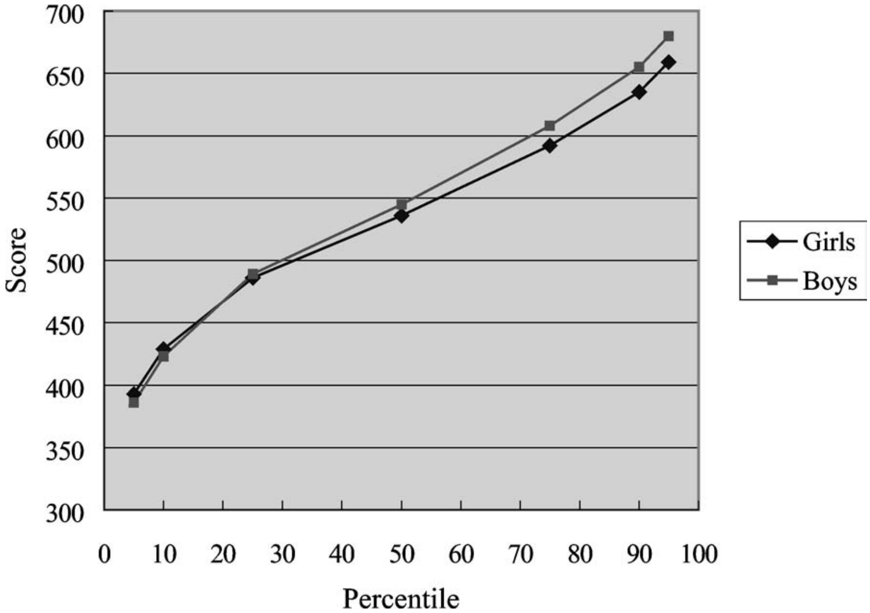

Figure 2. Hong Kong girls' and boys' scores in scientific literacy.

more persistent and pronounced gender difference in university intake for science courses.

The pattern of gender difference in performance across different ability groups of Hong Kong students can be seen more clearly in Figure 2. At the lower levels, boys and girls do not significantly differ, but the difference

TABLE III

Hong Kong girls' and boys' scores on different scientific literacy skills

Ability	Gender	Mean score	Standard error
1. Understanding concepts	Girls	236.94**	0.8
	Boys	240.79**	0.8
2. Recognising questions	Girls	76.03*	0.2
	Boys	75.30*	0.2
3. Identifying evidence	Girls	79.40*	0.2
	Boys	78.74*	0.2
4. Drawing conclusions	Girls	98.53	0.3
	Boys	98.64	0.3
5. Communicating conclusions	Girls	48.14	0.1
	Boys	48.30	0.1
Processes of scientific inquiry	Girls	253.96	0.6
[Abilities 2, 3 and 4]	Boys	252.68	0.6

*Difference of the means is significant at the 0.05 level.
**Difference of the means is significant at the 0.001 level.

increases with rising ability, and becomes quite distinct in the uppermost percentiles.

Gender Differences in Performance in Various Components of Scientific Literacy

Another way to analyse gender difference in science achievement is to compare the performances of girls and boys in different components of the scientific literacy framework (Table III). According to the results in Table III, the differences in performance between girls and boys are not statistically significant in 'drawing conclusions' and 'communicating conclusions'. However, the girls perform better than the boys in 'recognising questions' and 'identifying evidence', but less satisfactorily in 'understanding scientific concepts'. There is no statistically significant gender difference in the combined scores on the processes of scientific inquiry that involve 'recognising questions', 'identifying evidence' and 'drawing conclusions', and in the total scientific literacy scores.

The above analysis indicates that girls and boys show differential performance in different components of scientific literacy. This observation cautions us to be careful when drawing implications about gender differences in science achievement based on students' performance in assessment tests. A test dominated by items that assess understanding of

TABLE IV

Performance of Hong Kong girls and boys on closed and open-response items

	Performance (%)	
	Girls	Boys
Closed items	54.3*	57.3*
Open-response items	51.5	52.2

*Difference of the means is significant at the 0.01 level.

scientific knowledge may favour the boys, whereas a test that assesses both understanding of scientific knowledge and processes of scientific inquiry in a balanced proportion may lead to a different conclusion about the gender effect on scientific literacy. In the PISA 2000 study, the five components of the scientific literacy framework constitute different weightings towards the total score, with 42% of the total score on 'understanding scientific knowledge', 13% on 'recognising investigable questions', 13% on 'identifying evidence', 18% on 'drawing and evaluating conclusions', and 13% on 'communicating conclusions' (Table I). The inclusion of various components of the scientific literacy framework in PISA is quite different from the greater emphasis of TIMSS on understanding of scientific concepts. As PISA assesses a whole spectrum of abilities in scientific literacy, the advantage of boys over girls in items on scientific knowledge may have been offset by their weaker performance in items concerned with certain processes of scientific inquiry. This may help to explain in part why there is no significant gender difference in performance on the PISA science test for the student population of Hong Kong, whereas the boys of Hong Kong show much better performance than the girls in the TIMSS science assessment.

Gender Differences in Performance on Closed and Open-Response Items

A number of studies suggest that boys out-perform the girls on multiple-choice items while girls perform better in open-response items (e.g., Bolger & Kellaghan, 1990; Volkoff & Hocevar, 1995). The PISA assessment instrument for scientific literacy is made up of both closed and open-response items, which contribute to about 60% and 40% of the total scores respectively. To explore the possible gender effect of such items on science performance in PISA, the performances of the girls and boys of the Hong Kong sample on both types of items are compared in Table IV.

TABLE V

Performance of Hong Kong girls and boys on items in different thematic areas

	Performance (%)	
	Girls	Boys
Earth and environment	47.3*	51.1*
Life and health	56.8	56.7
Technology	59.0**	61.1**

*Difference of the means is significant at the 0.01 level.
**Difference of the means is significant at the 0.001 level.

The comparison shows that boys have an advantage over the girls on closed items, which include multiple-choice questions, true-false questions and questions that can be answered with single words or simple phrases. There is, however, no statistically significant gender difference in performance on the open-response items.

Gender Differences in Performance on Different Thematic Areas

It is generally believed that boys have an advantage over girls on assessment items set in physical science but not in biological science. To test the validity of this assertion, the performances of the boys and girls of the Hong Kong sample in items set in the three thematic areas of the PISA scientific literacy framework are compared in Table V.

The results indicate that boys perform better than the girls in items set in *Earth and environment* and *Technology*, which are in general related to physical science, while the performances of boys and girls on items set in *Life and health* are comparable. These findings are consistent with the observations made in other studies that boys tend to do better than girls in physical science but not in biological science (e.g., OECD, 2001, p. 126; Jovanovic et al., 1994).

Conclusions and Implications

The differences in performance on scientific literacy between the girls and boys of Hong Kong in PISA 2000 have implications for assessment and education policy while raising questions about schools and societal needs. Boys tended to score higher on ‘earth and environment’ and ‘technology’ items, ‘understanding of scientific knowledge’ items, and closed items.

Meanwhile, girls tended to score higher on 'recognizing questions' and 'identifying evidence' items. Also, boys scored higher than girls at the upper percentiles though their scores did not differ significantly for the overall population.

Student assessment implications can be seen through a comparison of TIMSS and PISA results. Compared to TIMSS, PISA has a more balanced combination of items from different thematic areas, more open-response items, less emphasis on understanding of science concepts but greater emphasis on processes of scientific inquiry and communicating conclusions. These differences may help to explain the overall non-significant gender difference in the scores of Hong Kong students in the PISA study.

The present observation is consistent with the findings from other studies that boys tend to out-perform girls in external standardised science tests, whereas girls often get higher science grades than boys in school-based assessment (Cole, 1997; Linn, 1991). Standardised tests usually are based on specific and limited samples of student work, have a greater weight on closed items and on scientific knowledge, whereas school-based grades are usually generated from more diverse types of exercises and activities assigned by the students' own teachers. The different patterns of performance of Hong Kong girls and boys in TIMSS and PISA provide evidence that varying the forms of assessment can produce different views of boys' and girls' academic competences. As both boys and girls seem to be disadvantaged by particular methods of assessment, gender bias in assessment may be reduced or eliminated by using science tests that assess diverse content and literacy skills in a variety of response formats.

But what constitute a gender-equitable assessment of scientific knowledge and skills? Consideration of the following questions can help science educators develop appropriate assessments of students' scientific literacy: *Are the weightings of test items assessing various knowledge and skills congruent with the aims of the curriculum and the needs of the society? Is the assessment made up of a balanced combination of items of different response formats?* Other deliberations include contexts of the test items, whether they are gender-biased, gender neutral or gender inclusive (e.g., Murphy, 1996; Rennie & Parker, 1993), and test-taking situations, as males tend to have an advantage in external, timed assessment situations while females have an advantage on school-based, cumulative assessment tasks (e.g., Hildebrand & Allard, 1993; Parker & Tims, 1994). Such considerations regarding the key features of fair and equitable assessment can provide a more holistic and comprehensive picture of the science achievement of students of different gender, backgrounds or learning styles, thereby promoting teaching of the desired scientific knowledge and skills for addressing a society's social, economic and political needs. Accurate assessment of students' strengths and weaknesses can inform educational policy so that the society can allocate its resources efficiently.

Gender-biased assessment also has important education and career implications. Traditional science tests, with focus on scientific knowledge but little emphasis on other scientific literacy skills, typically favour boys and portend continued male leadership in science-related fields. Such assessment adversely affects girls' science achievement and consequently their self-concepts in science, and may discourage or bar them from pursuing further study in science at higher levels (Oakes, 1990). As a result, the leaders in science-related fields tend to be men. Despite the efforts of test setters to be fair and equitable, the different interests of men and women may contribute to over-allocation of time, effort and other resources into areas that interest men, particularly in content related to physical science, and under-allocation of resources to issues that interest women. Moreover, a paucity of women in science-related fields reduces gender diversity and probably also reduces the diversity of knowledge and experiences that can inform both societal policy-making and implementation. The present-day gender gap in science achievement may be ameliorated by a greater awareness and concern for the need to replace traditional science achievement tests with assessment practices that are more gender equitable, and the PISA assessment instrument provides an example of how this can be achieved.

ACKNOWLEDGEMENTS

This study was supported by the Quality Education Fund of the Hong Kong Government. The opinions are those of the authors and not necessarily those of the funding agency.

REFERENCES

Blithe, T., Clark, M. & Forbes, S. (1994). *The testing of girls in mathematics.* Wellington: Victoria University.

Bolger, N. & Kellaghan, T. (1990). Method of measurement and gender differences in Scholastic Achievement. *Journal of Educational Measurement, 27*, 165–174.

Browne, N. & Ross, C. (1991). Girls' stuff, boys' stuff: Young children talking and playing. In N. Browne (Ed.), *Science and technology in the early years.* Buckingham: Open University Press.

Cole, N.S. (1997). Understanding gender differences and fair assessment in context. In W.W. Willingham and N.S. Cole (Eds.), *Gender and fair assessment* (pp. 157–183). London: Lawrence Erlbaum.

Gorard, S., Salisbury, J. & Rees, G. (1999). Reappraising the apparent underachievement of boys at school. *Gender and Education, 11*(3).

Gorman, T.P., White, J., Brook, G., Maclure, M. & Kispal, A. (1988). *Language performance in schools: Review of APU language monitoring 1979–1983.* London: HMSO.

Harding, J. (1979). Sex differences in examination performance at 16+. *Physics Education, 14*, 280–284.

Head, J. (1999). *Understanding the boys: Issues of behaviour and achievement.* London: Falmer Press.

Hildebrand, G. & Allard, A. (1993). Transforming the curriculum through changing assessment practices. In S. Haggerty and A. Holmes (Eds.), *Transforming the curriculum: Our future depends on it.* Waterloo: University of Waterloo.

HKPISA (2003). *The first Hong Kong PISA report.* Hong Kong: Hong Kong PISA Centre.

Hoste, R. (1982). Sex differences and similarities on performance in a CSE biology examination. *Educational Studies, 8*, 141–153.

Johnson, S. (1987). Gender differences in science: Parallels in interest, experience and performance. *International Journal of Science Education, 9*, 467–481.

Jovanovic, J., Solano-Flores, G. & Shavelson, R.J. (1994). Performance-based assessments: Will gender differences in science achievement be eliminated? *Education and Urban Society, 26*, 352–366.

Kahle, J.B. & Lakes, M.K. (1983). The myth of equality in science classrooms. *Journal of Research in Science Teaching, 20*, 131–140.

Keeves, J.P. (1986). Science education: The contribution of IEA research to a world perspective. In N.T. Postlethwaite (Ed.), *International educational research, papers in honor of Torsten Husen.* Oxford: Pergamon Press.

Keeves, J.P. (1992). *Learning science in a changing world, cross-national studies of science achievement, 1970 to 1984*. The Netherlands: International Association for the Evaluation of Educational Achievement.

Kimbell, R., Stables, K., Wheeler, T., Wosniak, A. & Kelly, V. (1991). *The assessment of performance in design and technology.* London: School Examinations and Assessment Authority.

Law, N. (1996). *Science and mathematics achievements at the junior secondary level in Hong Kong.* Hong Kong: TIMSS Hong Kong Study Centre.

Linn, M.C. (1991). Gender differences in educational achievement. In *Proceedings of the 1991 ETS Invitational Conference: Sex equity in educational opportunity, achievement, and testing* (pp. 11–50). Princeton, NJ: Educational Testing Service.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale: Erlbaum.

Mullis, I.V., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L. & Smith, T.A. (1998). *Mathematics and science achievement in the final year of secondary school.* Boston: Center for the Study of Testing, Evaluation and Educational Policy, Boston College.

Mullis, I.V., Martin, M.O., Fierros, E.G., Goldberg, A.L. & Stemler, S.E. (2000). *Gender differences in achievement: IEA's Third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Murphy, P. (1991). Gender differences in pupils' reactions to practical work. In B. Woolnough (Ed.), *Practical science*. Milton Keynes: Open University Press.

Murphy, P. (1996). Gender and assessment in science. In L.H. Parker, L.J. Rennie and B.J. Fraser (Eds.), *Gender, science and mathematics: Shortening the shadow* (pp. 105–117). Dordrecht: Kluwer Academic Publishers.

Murphy, P. (1997). Gender differences: Messages for science learning. In K. Harnquist and A. Bergen (Eds.), *Growing up with science: Developing early understanding of science*. London: Jessica Kingsley.

Murphy, P. (1999). Supporting collaborative learning: A gender dimension. In P. Murphy (Ed.), *Learners, learning and assessment.* London: Paul Chapman Publishing and Open University.

National Assessment of Educational Progress (NAEP) (1978). *Science achievement in the schools: A summary of results from the 1976–77 National assessment of Science*. Washington, DC: Education Commission of the States.

Oakes, J. (1990). *Lost talent: The underparticipation of women, minorities and disabled persons in science.* Santa Monica: RAND.

Organisation for Economic Co-operation and Development (OECD) (2001). *Knowledge and skills for life: First results form PISA 2000.* Paris: OECD.

OECD (2002). *Manual for the PISA 2000 database*. Paris: OECD.

OECD (2003). *Knowledge and skills for life: Second results form PISA 2000.* Paris: Organisation for Economic Co-operation and Development.

Parker, L.H. & Tims, J.E. (1994). Different modes of assessment in science and mathematics: A systematic interaction with gender. Paper presented at the annual meeting of the National Association for Research in Science Teaching, San Francisco, CA.

Rennie, L.J. & Parker, L.H. (1993). Assessment in physics: Further exploration of the implications of item context. *Australian Science Teachers Journal, 39*, 28–32.

Samejima, F. (1969). Estimation of latent ability using a response patter of graded scores. *Psychometric Monograph, No. 17.*

Schmidt, W.H., Raizen, S.A., Britton, E.D., Bianchi, L.J. & Wolfe, R.G. (1997). *Many visions, many aims: A cross-national investigation of curricular intentions in school science*. Dordrecht: Kluwer Academic Publishers.

Volkoff, J. & Hocevar, D. (1995). Multiple-choice versus performance-based testing and gender. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Warm, T.A. (1985). *Weighted maximum likelihood estimation of ability in Item Response Theory using tests of finite length.* Technical Report CGI-TR-85-08, U.S. Cost Guard Institute, Oklahoma City.

Whitehouse, H. & Sullivan, M. (1992). *Girls and year 12 science examinations.* Adelaide: Senior Secondary Assessment Board of South Australia, Adelaide, Australia.

Din Yan Yip
Department of Curriculum and Instruction,
The Chinese University of Hong Kong, Shatin,
Hong Kong
E-mail: yip2000@cuhk.edu.hk

Ming Ming Chiu
Department of Educational Psychology,
The Chinese University of Hong Kong,
Hong Kong

Esther Sui Chu Ho
Department of Educational Administration and Policy,
The Chinese University of Hong Kong,
Hong Kong