Claire Wyatt-Smith
Joy Cumming
Editors

Educational Assessment in the 21st Century

Connecting Theory and Practice

Springer

Chapter 7
Assessment, Gender and In/Equity

Susan M. Brookhart

Introduction

This chapter examines how males and females perform on different types of assessment tasks and in different disciplines. The focus is on assessment tasks that indicate students' levels of achievement in the academic disciplines taught in school: reading/language arts, mathematics, science and social studies. The population of interest is school-aged children. Three sections develop a line of inquiry based on questions. First, are there gender differences in achievement? If so, what might they mean? And does the assessment process itself contribute to creating them? The approach taken here is to examine large-scale national and international studies of achievement, using, when possible, standardised measures of effect sizes. This book presents international perspectives on student achievement, and thus this chapter aims to report gender issues across national borders. It relies on studies where achievement outcomes were measured with different assessments. Standardised measures are required in order to make valid comparisons from country to country and from assessment to assessment. Classroom processes, including classroom assessment, are the most important aspects of schooling, and the classroom is the source of achievement measured by standardised tests. However, studies of classroom assessments were not used in this review because of the chapter's purpose. The theoretical and methodological discussions at the end of the chapter describe in detail the methodological choices.

Are There Gender Differences in Achievement?

In a discussion of assessment and gender, the first and obvious question that must be dealt with is: Does student achievement differ by gender? If the answer to that question is no, then follow-up questions become moot. If the answer is yes, then it is

S.M. Brookhart (✉)
Duquesne University, Helena, MT, USA
e-mail: susanbrookhart@bresnan.net

C. Wyatt-Smith, J.J. Cumming (eds.), *Educational Assessment in the 21st Century*,
DOI 10.1007/978-1-4020-9964-9_7,

important to describe the differences and ask how educators and others have interpreted their meaning. This chapter begins with a brief review of studies investigating gender differences in achievement on standardised tests.

Reading and Language Arts

The Organisation for Economic Co-operation and Development (OECD) began its Programme for International Student Assessment (PISA) in 1997, in an effort to collect internationally comparable information about student performance and related student, family and institutional factors that could inform policy making. The first PISA survey assessment was conducted in 2000. There was another PISA survey in 2003, and one in 2006. The 2006 survey included data from 30 OECD member countries and 27 partner countries.

Gender differences in reading have been evident in all three PISA surveys (OECD, 2007). In the 2006 PISA survey, gender differences in reading, favouring females, existed in every country. These differences were between 20 and 57 points, averaging 38 points. The overall standard deviation in reading is 99, so the effect sizes of these differences are between 0.20 and 0.58, averaging 0.38, which puts them in the 'small-' to 'medium-effect-size' range (Cohen, 1988). The difference was found in every country surveyed. The 2006 PISA reading data from the United States were not used in the analysis because of an error in printing the reading test booklets, so there were 56 countries with reading data and 56 countries with a gender gap in reading, favouring girls.

In the United States, the National Assessment of Educational Progress (NAEP) is the only nationally representative, ongoing study of student achievement. NAEP measures reading comprehension by asking students to read passages and answer questions about what they have read. NAEP data, like the PISA data internationally, find a consistent gender gap in reading, favouring girls. Klecker (2006) analysed NAEP reading data for the public school samples from 1992, 1994, 1998, 2002 and 2003. Effect sizes were small in 4th grade (0.13–0.27) and small to moderate in 8th grade (0.27–0.43) and 12th grade (0.22–0.44). The 2007 NAEP data (Institute of Education Sciences, 2007) also show females outscoring males in reading by 7 points (effect size of approximately 0.20) in 4th grade and by 10 points (effect size of approximately 0.29) in 8th grade, continuing the same pattern.

Lietz (2006) studied gender differences in reading across English and non-English speaking countries. Her stated research purpose was to use modern statistical techniques (meta-analysis and hierarchical linear modelling) to address the question of gender differences in reading in order to address conflicting reports in the literature, much of which reported that girls out-performed boys in reading, but some of which did not. Her meta-analysis included 139 effect sizes from various studies of secondary school reading achievement between 1970 and 2003, including the International Association for the Evaluation of Educational Achievement (IEA) Reading Comprehension Study (1970–1971) and Reading Literacy Study (1990–1991), PISA 2000, NAEP 1992–2003, a number of national studies in Australia

over the period 1975–2002 and other published studies. The overall grand mean was an effect size of 0.19, a small effect that meant girls outscored boys overall. Gender differences were most pronounced in PISA, followed by the Australian assessment programs and NAEP. The effect for gender held whether English was the language of test administration or not, and the effect for gender did not increase or decrease with age. There was also some unexplained variance, which meant the predictors used did not completely explain differences in effect sizes among the studies.

Mathematics

The literature on gender differences in mathematics is more variable than findings about gender differences in reading and language arts. Many authors report that boys perform better than girls in mathematics, and cite some literature to support that in literature reviews preceding their own studies of the matter. However, differences between the genders on mathematics achievement are small, when they do exist, are not consistent among countries and often are washed out with between-country differences. Thus, the cultural argument—that differences in performance when they exist are most likely due to differences in curriculum, instruction, opportunity to learn, and cultural, political or social factors—is easily supported for mathematics.

In the 2006 PISA survey in mathematics, there were smaller gender differences than in reading, favouring males. Males outperformed females in some countries, but not by much, with an average 11-point difference (effect size 0.12). In Qatar, however, females outperformed males in mathematics (OECD, 2007). Based on previous PISA research, the OECD used a cultural and economic argument to explain the differences, most notably that the tendency of males to outscore females in mathematics is mitigated by the tendency for females to attend higher performing school programs (OECD, 2007, p. 324).

This conclusion (about the effect of school program) is reinforced by a comparison made in PISA 2003, when PISA also measured student performance in problem solving, reported in *Problem Solving For Tomorrow's World: First Measures of Cross-curricular Competencies* from PISA 2003 (OECD, 2004). This suggested that males and females perform roughly equally in analytical reasoning skills, which also form one component of mathematics tasks. The gender differences in mathematics appeared to correspond to the contexts in which tasks are embedded at school, rather than to the underlying mathematical reasoning skills.

Beller and Gafni (1996) compared the performance of 9- and 13-year olds in 14 and 20 countries, respectively, on the mathematics and science portions of the second International Assessment of Educational Progress, with data collected in 1991. Consistent with other studies reviewed, they found effect sizes (corrected for attenuation due to unreliability) of 0.05 and 0.12 for 9- and 13-year olds, respectively, for mathematics overall performance and 0.17 and 0.30 for 9- and 13-year olds, respectively, for science overall performance. The same trends were found within countries, but with some differences in the magnitude of the differences.

Ethington (1990) analysed data from the 1981–1982 Second International Mathematics Study (SIMS). She used median polishing, an exploratory data analysis method that does not require a priori hypotheses. Largest differences in medians were associated with country. The effects of gender were small and went both ways. The largest gender effect favouring females was an expectation of 1.5 percentage point better score on fractions, and the largest effect favouring males was an expectation of 1.5 percentage point better score on geometry. However, there were interaction effects between country and gender; for example, there was more of an effect against girls in France and in favour of girls in Thailand.

Science

In addition to reading and mathematics, PISA 2006 also studied science performance. In the combined science scale, most of the 57 countries (30 OECD and 27 partner countries) had no significant differences. Among the 14 that did show significant differences, only four of those differences had an effect size of 0.20 (a 'small' effect) or greater, three countries favouring females (Jordan, Qatar and Thailand) and one favouring males (Chile). Similarly, there were not many countries with gender differences on a 'general interest in science' scale, and those differences were mixed (some favouring boys and some girls). OECD (2007, p. 163) concluded that there were no entrenched performance differences by gender, but there were gender differences in attitudes towards science, and these differences varied by country.

For example, in many countries there were differences favouring girls on scales measuring 'identifying scientific issues' and 'level of concern for environmental issues' and differences favouring boys on 'self-confidence in science'. Other scales measuring science attitudes produced mixed results for gender. Looking across attitude scales, PISA 2006 identified countries in which males had higher average scores on at least five attitude scales: Germany, Iceland, Japan, Korea, the Netherlands, the United Kingdom, Chinese Taipei, Hong Kong-China and Macao-China—although in Iceland, Germany and the Netherlands, females also had at least one higher-than-average attitude scale ('concern for environmental issues' or 'responsibility for sustainable development') than males, as well. Even though there were few differences in science achievement, gender differences in science attitudes were of concern because of their potential effects on future education and career choices.

Multiple Subjects

Some studies have looked at gender differences across subjects. This section reviews results from just a few. The selection of studies is intended to be illustrative, not exhaustive. Like the PISA project, these studies have assessed achievement in school subjects by school-aged students.

Nowell and Hedges (1998) looked at seven surveys of 12th-grade student achievement in the United States from 1960 through to 1992, plus the 1971–1994 National

Assessment of Educational Progress, in multiple subjects (reading, mathematics, science and writing in all surveys including NAEP, plus vocabulary and perceptual speed subtests in some of the other surveys). An important feature of their study is that they examined differences in means, variances and extreme scores. Differences in means and variances were small, while differences in extreme scores were sometimes substantial. Writing produced the largest mean differences (moderate effect sizes of 0.48–0.61, favouring females). Females also outperformed males in reading (small effect sizes of 0.00–0.30). Males outperformed females, on average, in science (small-to-moderate effect sizes, 0.22–0.51) and mathematics (mostly small effects, and a few moderate, 0.09–0.40).

Nowell and Hedges (1998) also did a trend analysis and concluded that the gender differences have not changed significantly over time, with the exception of NAEP science scores and non-NAEP mathematics scores, both in the direction of closing the gender gap somewhat. Nowell and Hedges also analysed variances; in almost all samples, males' performance was more variable than females. They also analysed the proportions of males and females in the extremes of score distributions and found that males were over-represented in the upper tails of score distributions for mathematics, science and survey composite scales. Females were over-represented in the upper ends of reading, vocabulary and perceptual speed, but to a lesser degree than was the case for males in mathematics, science and composites.

Sammons (1995) used hierarchical linear modelling to study gender, ethnic and socio-economic differences in student achievement on Britain's 'School Matters' student cohort, which followed students over the period 1980–1984 as they progressed through public examinations in years 3, 5 and 6 (transfer to secondary school), and then checked their performance on the General Certificate of Secondary Education in 1989 (GCSE, for year 11 students). The statistical modelling technique allowed the effects of background characteristics to be estimated as net effects, with other background and school membership characteristics controlled. Gender differences (favouring girls) were found in reading but not in mathematics in year 3, and in both reading and mathematics (favouring girls in both subjects, although the mathematics effect was smaller) for year 5, and gender helped predict mathematics progress (but not reading progress) between these years. However, socio-economic effects were larger than gender effects. By the time of GCSE, in year 11, gender effects still favoured girls but socio-economic effects were still stronger than gender effects. Sammons (1995) noted that the junior school tests were not multiple choice and that the GCSE included course work.

DeLisle, Smith, and Jules (2005) studied primary school achievement as measured by the 2004 national examinations for standard 1 (7–8-year olds) and standard 3 (9–10-year olds) and by the 2003 Secondary Entrance Assessment (11–12-year olds) in Trinidad and Tobago. They found that girls had an advantage across all assessments at different grade levels on both language and mathematics, but the gap was very small in mathematics—not of practical significance—and decreased at higher-grade levels. Girls had larger advantages in language arts and creative writing, some of which DeLisle et al. (2005) judged to be practically significant. Girls'

advantage in language arts was most evident for pupils in lower-ability groups, higher-grade levels and rural educational districts. In answer to their title question, 'Which males and females are most at risk and on what?' therefore, they pointed out that not all boys but rather boys who were doing poorly at school overall and who lived in rural areas in Trinidad and Tobago were the at-risk students.

In the United States, many high school students take one of two voluntary college admission tests (the ACT® and the SAT®) to include with their college applications. While this results in a self-selected sample, interest in potential gender differences on these tests runs high because of the potential implications for students' college admission and future study.

The ACT® test covers English, mathematics, reading and science, and reports a composite score. For the ACT®, the popular belief had been that boys outscored girls (male ACT® composite average was about 0.2 points higher than females for 1999–2004; ACT, 2005). ACT's (2005) analysis of two states that had adopted the ACT as part of required state assessment for all students showed that the male score advantage was a result of self-selection. The gap changed directions (to a 0.1- and 0.2-point advantage for the girls in the two states, respectively) for the states in which all students took the ACT®. These score differentials are tiny and, ACT (2005) concluded, not of practical significance.

Coley (2001) wrote an analysis of gender and racial/ethnic differences in achievement for Educational Testing Service, publisher of the SAT®. He found that black college-bound seniors were the only group in which girls scored higher than boys on the SAT® I Verbal Test. On the SAT® I Mathematics Test, boys in all racial/ethnic groups scored higher than girls. More girls than boys took Advanced Placement (AP) exams; these are challenging exams that can earn students course credit in college, typically by scoring a 3 or better on a 5-point scale. Among students who scored in this range, there were few differences in the percentage of males and females on the AP Literature and Composition Examination. However, a higher percentage of boys than girls scored in this range on the AP Biology or AP Calculus AB Examination.

Conclusions Across Subjects

The results of these studies suggest that the answer to the question, 'Are there gender differences in achievement?' is 'Yes and no'. There does appear to be a rather robust 'gender gap', favouring girls, in language arts in most countries, with effect sizes reported in the small-to-medium range. However, in other subjects, there may or may not be differences, depending on the country. Gender differences in achievement tend to be small in comparison with socio-economic differences, with racial/ethnic differences in some countries, and with differences between countries. Some countries, for example, have gender differences favouring boys in mathematics, and some do not.

A notable point from the PISA study is that patterns of variability are more dramatic than mean differences. There was more variation in gender gap size between

schools in any one country than between countries. Also, within-country variance was greater than between-country variance (OECD, 2007). This means that, whatever differences were apparent in aggregated data, individual boys and girls, and individual schools, may be very different from the average. This also means that, even where gender differences exist, gender explains at most a small part of student performance.

What Do Gender Differences in Achievement Mean?

Having established that there are gender differences at least in language arts achievement, the next logical question is what those differences mean. Is there a 'gender gap' sizable enough to be of practical significance? What causes this gap? Can results be altered by changes in policy or educational program?

What Are the Different Stances on This Issue?

Interpreting whether the gender gap is of practical significance, and, if so, what to do about it, reflects the interpreter's theoretical stance in philosophy, sociology, psychology, economics and politics. The two basic divisions are between 'nature' and 'nurture' stances, the former privileging biological differences as explanations and the latter privileging culture, upbringing, education and experience (Francis & Skelton, 2005).

What Is the Author's Bias with Regard to the Meaning of Gender Differences in Achievement?

'Nature' and 'nurture' are not mutually exclusive, and I do not believe that biological or cultural differences are necessarily 'bad', either. Various cultures have developed to make sense of the world and life in it. No true experimental manipulations can be done—researchers cannot 'assign' students to cultures or genders. However, a gender effect across cultures is more supportive of biologically based explanations, whereas variation in gender effects between cultures is more supportive of culture- and experience-based explanations.

For the purpose of interpreting assessment results, it is important to consider what constructs the assessments were designed to measure. I subscribe to the view that true learning implies the ability to use knowledge. Thus, I find evidence from the PISA assessments more persuasive than, for example, basic skills tests, because PISA took a 'literacy' approach to the constructs measured by the assessments. Reading literacy was defined in PISA as understanding and using written material in order to develop one's knowledge and potential. Similarly, PISA measured mathematical literacy as the ability to analyse, reason and communicate as they solve

mathematical problems. Scientific literacy, a focus for PISA 2006, was measured as the ability to understand and use science concepts and to think scientifically about evidence (OECD, 2007).

Two other points about my approach combine both assessment and methodological concerns. As an educator, I believe that relative comparisons ('Who outscored whom?') are less important than change over time ('What progress is being made?'). I also believe that relative comparisons are less important than descriptions of performance capabilities: the answer to the question 'Who is better, boys or girls?' is less important than the answer to 'What can boys and girls do now?' and 'What else could they be expected to do next?' Relative comparisons are not as useful for making instructional improvements as information about progress and performance.

These theoretical and methodological stances contribute to the discussion in this chapter. Another reader might draw somewhat different conclusions from the same studies.

How Can Knowledge of Gender Differences in Achievement Help Inform the Assessment Process?

Francis and Skelton (2005) pointed out that different countries have responded to the 'gender gap' news with different levels of alarm, and with different educational policies. Australia and the United Kingdom, for example, reacted with policy documents about gender equity that were concerned in particular with the 'underachievement' of boys. Australia, especially, has been noted for its strong policy documents arguing for the education of boys in 1997 (*Gender Equity: A framework for Australia's schools*) and 2002 (*Boys: Getting it right. Report into the Inquiry of Education of Boys*).

In the United States, however, the 2001 *No Child Left Behind* legislation has focused on equity among groups based on ethnicity, socio-economic status, student-disability status and English-proficiency status. Schools must report student achievement data disaggregated according to these groups, but not gender.

Some researchers have intentionally addressed issues of interpretation in their research questions, study designs and discussion of results. Robertson (2005) presented results of a series of international surveys and the Scottish Assessment of Achievement Programme (AAP) that showed a small but statistically significant gender gap in mathematics, favouring boys especially in some sub-domains. In 1988, girls were better at whole-number arithmetic but boys were better at measurement, area, and some other sub-domains, depending on age. By the early 1990s and continuing to 2000, differences had disappeared. Robertson (2005) interpreted this closing of the gap in terms of government and school policy changes.

Duffy, Gunther, and Walters (1997) examined gender differences in mathematical problem solving, and interpreted their results as supporting the socialisation or 'nurture' (as opposed to the biological or 'nature') explanation. They measured attitudes towards mathematics as well as problem solving. They found a complex relationship between gender and mathematics: there were no systematic gender differences on one test (GAUSS) but there were on another (CTBS), overall and among

the top 10 per cent in ability. Content experts, however, rated the GAUSS questions as being more abstract and difficult. Attitudes predicted performance on testing at one occasion but not at another. The authors reasoned that, if biological differences were the explanation for mathematics performance difference, there would have been gender differences in performance on both tests.

Another study took a developmental approach. DeFraine, Van Damme, and Onghena (2007) studied the relationship between academic self-concept and achievement in Dutch, in Flemish students (Flanders is the Dutch-speaking part of Belgium). Changes in self-concept and achievement were not related, although there was a positive relation between self-concept and achievement. In secondary school self-concept declined, faster for girls than boys, and achievement rose for girls but dipped and then rose for boys. Achievement was high overall in Flemish-speaking schools in Flanders. These developmental changes are more congruent with a sociological than a biological explanation because they are situated in students' educational experiences.

What Is Known About Gender Differences in Assessment Development?

This section focuses on whether assessments themselves—differences in assessment design, development, administration and use—could be responsible for observed gender differences. Some studies have attempted to address aspects of assessment hypothesised to explain part of the gender gap. These have mostly focused on questions of assessment format, usually with the hypothesis that girls will do better on performance assessment and problem-solving tasks (variously theorised to be because of their more interactive nature or because of their language components) than on traditionally formatted tests and basic skills questions. Willingham and Cole (1997) and their contributing authors published a landmark review of these issues. The literature they reviewed did not make a compelling case that any of the assessment aspects studied provided major explanations for gender differences. The results of more recent studies have not changed that conclusion.

Efforts to Remove Gender Bias During Assessment Development

Professional standards for test developers require that they try to prevent differences by gender and any other categories that should be irrelevant to the construct to be measured. For example, in the United States the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2004) states that test developers should 'obtain and provide evidence on the performance of test takers of diverse subgroups, making significant efforts to obtain sample sizes that are adequate for subgroup analyses. Evaluate the evidence to ensure that differences in performance are related to the skills being assessed'.

In the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), Section "Test Administration and Test-Takers' Behaviour" is devoted to 'Fairness in Testing and Test Use'. The introduction to the section points out that there are many definitions of 'fairness' in testing, including lack of bias, equitable treatment in test administration, equality of outcomes and opportunity to learn. Bias, on the other hand, refers to 'construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees' (p. 76). To use an extreme hypothetical example to illustrate, if in some country girls were allowed to go to school but boys were not, and gender differences in achievement were noted, that would not be 'fair' in the sense that girls and boys did not have equal opportunity to learn, but it would not necessarily mean the test was biased. The test might be measuring real differences in achievement in a valid manner.

At the item-development stage in test construction, sensitivity reviews typically have panels of experts to review item content. Items that are offensive or that conceptually seem to favour one group over another are edited or discarded. At the pilot testing stage, empirical data are reviewed for evidence of construct-irrelevant bias. The term 'differential item functioning' (DIF) analysis is used for methods that statistically compare performance between reference and focal groups of students. DIF methods hold achievement constant; for instance, a gender DIF analysis would compare performance on a given item between boys and girls at the same achievement level. Many statistical methods are available to study DIF. There are methods by which to evaluate DIF for both multiple-choice or other right/wrong items and for multi-point items or tasks. DIF analysis is used routinely in the preparation of large-scale tests. Sometimes, validity evidence offered for tests also includes studies designed to test for differential prediction; for example, if a test score was a better predictor of grades or other future achievement scores for girls than for boys.

Therefore, the reader should not expect this section to find major assessment effects that are large enough to explain gender differences in achievement. Standard procedures require reviews, conceptually at the item development and review stage and empirically at the pilot-testing stage, that are designed to prevent construct-irrelevant gender differences in assessment items or tasks.

The following sections summarise what is known about gender differences in various aspects of the assessment process. Readers who would like more detail should consult Willingham and Cole's (1997) book-length review of this topic.

Choice of Construct and Test Content

Content of multiple-choice items. Can the content of multiple-choice items affect girls' and boys' performance differently? Bridgeman and Schmitt (1997) reviewed exploratory studies of gender DIF conducted before such analyses became a routine part of test development. They concluded that these early DIF studies supported the finding that items with content about human relationships or aesthetics/philosophy

were differentially easier for females and that items with science-related content and specialised terminology were differentially easier for males. In mathematics, algebra items were differentially easier for females and geometry items were differentially easier for males. Abstract, pure mathematics problems were differentially easier for females, and mathematics word problems were differentially easier for males. As an example of the use of specialised terminology, Bridgeman and Schmitt (1997) used an item from one of the studies they reviewed (Curley & Schmitt, 1993, cited in Bridgeman & Schmitt, 1997). The analogy 'vortex:water::' with the correct answer choice 'tornado:air' was differentially more difficult for females. Changing the item stem to 'whirlpool:water::' removed the gender DIF and also made the item easier overall.

Because current practice specifies the routine use of DIF analyses in test development, DIF analyses conducted now with operational tests would not find much in the way of gender differences. Potentially biased items are discarded before test forms are finalised. In addition, analyses at the item level are instructive, but not the whole picture. Total test (or subtest) scores rely on sets of items that represent the construct of interest in a balanced manner.

Content of essay prompts. Perhaps test items that require students to write could have more of an effect than multiple-choice items. Can the content of essay prompts affect girls' and boys' performance differently? Breland and Lee (2007) studied gender differences in scores from the computer-based Test of English as a Foreign Language (TOEFL®-CBT) essay prompts from 1998 to 2000. A total of 87 different essay prompts were studied. While many of the prompts had significant gender DIF, none of the effect sizes were large enough to be characterised as having an important group effect. The mean effect size across prompts was –0.13, favouring girls.

Bridgeman and Schmitt (1997) studied gender differences in performance on the 1993 and 1994 Advanced Placement (AP) U.S. History and Biology Examinations in the United States. They found mean gender differences on some but not all of the essay questions (absolute value of effect sizes ranged from 0.00 to 0.21), although the differences were smaller than for the multiple-choice questions on the same exams. In U.S. History, what differences there were favoured girls in 1993 and boys in 1994, while in biology, they favoured boys in both years. Identifying the content of specific essay prompts that contributed to gender differences in AP examination performance is difficult, however, because students are allowed some choice about which essay they answer. Unlike DIF analyses, analyses of mean differences do not control for the achievement level of the students on the construct of interest, so if the essay prompts were each answered by somewhat different students, comparison is problematic. Bridgeman and Schmitt (1997) concluded by pointing out that in this situation, the optimal solution is to make sure to offer balanced choices for students.

Types of Assessment Items and Tasks

Can the format of an assessment affect girls' and boys' performance differently? This question has been studied in several ways. Some researchers have examined

gender differences between multiple-choice and constructed-response test items. Others have examined gender differences between tests and performance assessments.

Multiple-choice versus constructed-response items. Willingham and Cole (1997) reviewed studies testing for format effects between multiple-choice and constructed-response test items. Constructed-response test items ask students to formulate their own answer to a test question; multiple-choice items ask students to select an answer from a list of options. Neither of these formats is an extended-performance assessment task. Willingham and Cole (1997) reviewed 12 studies that looked at whether multiple-choice and constructed-response test items measured the same construct and whether format differences were associated with gender differences. None of the studies that looked for gender differences by format found them, and most of the studies that looked for construct differences did not find them, either. They concluded (Willingham & Cole, 1997, p. 276) that available evidence did not support the hypothesis that multiple-choice format per se was a significant source of gender differences in test results.

More recently, DeMars (1998) studied performance on the mathematics and science sections of the state of Michigan's High School Proficiency Test in the United States. DeMars was interested in whether gender differences would appear between the two item formats used on the test: multiple-choice and constructed-response. There was no gender-by-format interaction on the mathematics test, and only a small interaction in science. When scores from only the top five per cent of students were used for analysis, males scored higher on the multiple-choice sections and females on the constructed-response sections (except for one form of the mathematics test, males still outscored females on the constructed-response section, but by a smaller margin than for the multiple-choice section). However, these differences were small, and in summary, DeMars (1998) did not judge any of the differences found to be of practical significance.

Tests versus performance assessments. If format-of-test items do not contribute to gender differences, what about tests versus more extended performance assessments? Performance assessments require observation and judgment of students' processes as they do tasks, and/or observation of products students create. Performance assessments take place over extended periods of time, are often complex and employ a degree of student choice.

Cox (2000) studied gender and urban/rural differences on student performance in the 1992 dataset of Common Assessment Tasks (CATs) for the Victorian Certificate of Education in Australia. There were four CATs: two were long-term performance assessments and two were more traditional examinations, one multiple-choice and one short-answer. Each CAT measured six subjects within mathematics curriculum for year 12 students (for example, space and number). For most subjects, girls outperformed boys on the two long-term performance assessments, and boys outperformed girls on the examinations. The hypothesis that girls are better with language was advanced as an explanation. The performance tasks required written work. In addition, the performance tasks allowed for drafts to be shared with teachers, and the authors hypothesised that girls might be more willing to submit to feedback and

pay attention to it. For the authors of the study, the urban/rural effect was equally troubling (city students had an advantage over rural students for many of the subjects and tasks).

Woodfield, Earl-Novell, and Solomon (2005) studied students at Sussex University in the United Kingdom, from first-year and last-year students in cohorts graduating in 1999, 2000 and 2001. They compared scores on two modes (course work versus examination) of assessment data from course work in a variety of different disciplines, in a mixed model analysis of variance (ANOVA) with gender as the between-groups factor. Results indicated that female students did outperform male students, but by about the same amount on both course-work assignments and examinations. The population of interest in this study was undergraduate students, not school-aged children. However, year 12 students and first-year undergraduates are very close in age. Woodfield and her colleagues' results are therefore of interest here because they used a sample from another country and obtained different results than did the Cox (2000) study.

Test Administration and Test-Takers' Behaviour

Bridgeman and Schmitt (1997) reviewed studies in several categories related to the conditions of testing and the behavioural responses of test takers. Time, time pressures and speededness of tests have been studied with respect to gender differences. Their review of these studies led Bridgeman and Schmitt (1997, p. 206) to conclude that 'at least on academic reasoning tests, time limits do not appear to be an important consideration in explaining gender differences in test scores'. Studies of student guessing and omitting answers similarly failed to find gender-related effects. In addition, studies of the effects of changing answers failed to find differential gender effects on performance. That is, there were no differences between boys and girls in score gains or losses resulting from changing answers. Thus, it does not appear that gender differences in achievement are explained by test administration factors or differential test-taking behaviours.

Test anxiety is another area in which gender effects have been hypothesised. Hembree (1988) did a meta-analysis of 562 studies of test anxiety and its relationship to performance. He found that test anxiety and performance were significantly related at Grade 3 and above. Across grades, girls exhibited higher mean test anxiety than boys, but their higher test anxiety did not appear to translate into a difference in performance. Again, cultural transmission seems to fit better as a hypothesis for test anxiety effects than a biological explanation.

Scoring of Items

Rater effects. Do rater effects explain gender differences in achievement on performance tasks? Bridgeman and Schmitt (1997) pointed out that while machine-scoring is blind to student characteristics, there could be rater effects associated with gender

or other student characteristics for hand-scored responses. In fact, this has been a classic question in the literature, based on a study (Goldberg, 1968) that suggested female raters were prejudiced against female students. Swim, Borgida, Maruyama, and Myers (1989) did a meta-analysis of 119 studies of gender effects among raters to investigate this stereotype. They found little evidence that raters evaluate females differently than they do males. Most of the mean effects they tested were not significantly different from zero, and where effects were found the effect size was very small. There was no main effect overall, for example, for gender of person rated; however, studies that did find differences were more likely to find those differences favoured males. Similarly, there was no interaction effect of gender of rater by gender of person rated; however, male raters exhibited more variability in ratings than did female raters.

Rubrics. Performance assessments are often scored with rubrics, which assign performance-quality levels under various criteria. The performance levels defined for rubrics should be written to index levels of achievement on the construct that a particular performance assessment is designed to tap. Therefore, asking whether the content of rubrics explains gender differences in achievement is not, strictly speaking, asking a question about construct-irrelevant variance. Nevertheless, at least one study did just that.

Wang and Lane (1996) studied differential item functioning on 33 mathematical thinking and reasoning items on the QUASAR Cognitive Assessment Instrument (QCAI), in the United States. Only two items were of concern. On one, girls with low mean total-test scores outperformed boys matched for total test score. On the second, girls with high mean total-test scores performed less well than boys matched for total-test score. The authors speculated that the second item's DIF may have appeared because it was on the same test form as the first. The first item read (Wang & Lane, 1996, p. 193): 'Jerome, Elliott, and Arturo took turns driving home from a trip. Arturo drove 80 miles more than Elliott. Elliott drove twice as many as Jerome. Jerome drove 50 miles'. The task was to write three mathematical questions that could be answered by this scenario, and/or with additional information provided in the student's response. The scoring rubric was based on the number, not the complexity, of the questions. Post-hoc analysis showed that girls wrote more questions than boys. Eighty-two per cent of girls wrote at least one question, compared with 76 per cent of boys. However, 60 per cent of boys, compared with 54 per cent of girls, wrote more complex questions. Therefore, a scoring rubric that took into account the complexity of the questions students wrote might have resulted in no differential item functioning by gender. The rubric encodes the intended construct into the score levels; the decision whether to score complexity as well as number should be based on the definition of problem solving used to represent the construct.

Summary and Conclusion

This chapter begins with the question as to whether gender differences in achievement do exist and, if so, where. There do appear to be gender differences in achievement in language arts that, while variable across countries and cultures, do

favour girls. Differences in mathematics and science are more variable. Where they exist, they are more likely to favour boys, but not always. This chapter presents evidence for those differences, based on the most recent PISA international comparison study and supported with corroborating findings from other studies.

The chapter then explores two different questions about the meaning of gender differences. First, the more general question of meaning, namely 'Are gender differences biologically or culturally based?' It appears that explanations for gender inequities that exist in some places are found in studies of culture, economics/politics and environment. Second, a more specific question addressed whether aspects of the assessment process itself might explain gender differences in achievement. Assessments seem to be able to indicate the issue, but in the main they are not the reason for differential performance by gender. While the conclusions that are drawn from an assessment should be informed by what is known about gender differences, it appears that what is known so far supports sociological, as opposed to biological or measurement-based, causal hypotheses.

Future research, then, will ask questions about achievement gaps in various countries, cultures and educational systems. Because gender equity at the item level will continue to be a quality-control issue in standardised test development, future research will look for explanations of gender differences in social, cultural and educational influences. Studies of policies that have brought equity—their operations, effects and unintended consequences—will join more theoretical studies to attempt to explain not only cause, but also solution strategies, for gender equity in achievement.

Challenges this future research will encounter include, first and foremost, the chicken-and-egg nature of questions about causes and influences of gender differences in achievement. Are differences a result of cultural and educational patterns or a cause of them, or both? Another challenge for future research is the relative importance of gender differences, which are mostly small, compared with economic and cultural differences in achievement. Given limited resources, studying gender differences might (and maybe should) give ground to studies of economic and cultural patterns in achievement, which may be more amenable to change.

Theoretical and Methodological Framings

The Quantitative Approach Taken in This Chapter

Studies used as evidence to answer this chapter's question 'Are there gender differences in achievement?' were large-scale studies, implementing national or international comparisons, using standardised tests. Effect sizes were the preferred statistics for reporting, where available. Following is the rationale for these choices.

Since the book presents international perspectives on student achievement, this chapter discusses gender issues in achievement across national borders. Standardised tests are less context-bound than classroom assessments or tests developed by researchers for particular evaluations or studies. While standardised test results

depend on student opportunity to learn, they also depend on opportunity to learn in the general sense, not in the specific sense of a classroom test, where opportunity to demonstrate knowledge of particular concepts and skills taught in the short term are more important. Standardised tests usually measure large-grain constructs like 'reading comprehension' or 'mathematics problem solving', rather than the ability to do one certain kind of reading or mathematics, as taught by one particular curriculum or group of teachers.

Effect sizes report research results in standardised terms. Because the aim is to answer a question about whether gender differences exist, it is important to standardise the comparisons between male and female students. Any differences reported should not be an artefact of the scale for the particular test given, or the number of students in the sample (as long as sample sizes were reasonable), or the number of items on the test, and so on. The 'effect size' used in this chapter is the standardised mean difference, sometimes called 'Cohen's *d*'. It is sometimes defined as the difference between mean performance of an experimental and control group, divided by the standard deviation of the control group. In this way it reports group differences in standard deviation units, which allows comparisons about the size of group differences from study to study, no matter what scale the outcome measure (in this case, an achievement test) used. In this chapter, differences between male and female students, as opposed to experimental and control group means, are compared, using usually a pooled standard deviation, to allow comparisons of gender differences across studies. For example, if girls outscored boys in language arts by 0.15 standard deviations on the achievement test used in one study and by 0.30 standard deviations on the achievement test used in another study, it is proper to conclude that the first difference is small and the second difference is moderately sized and larger than the first. These methodological choices are made to remove issues of tests and scaling as much as possible from the discussion, to allow concentration on the question of gender differences.

Glossary

Constructed-response format items Test questions for which the student responds with their own ideas (writing, drawing, working problems) instead of selecting from among prescribed choices

Differential item functioning (DIF) analysis Study of whether examinees of the same ability, but from two different groups, perform differently on a test item

Effect size While an effect size can be any of several standardised measures of the size of a result, in this chapter the effect size used is the difference between two groups' performance on an assessment expressed in standard deviation units.

Hierarchical linear modelling A method of analysis that takes into account the nested nature of data (for example, students within classrooms and classrooms within schools)

Median polishing A method of analysis to examine differences among factors, similar to analysis of variance but comparing medians for each factor rather than means

Meta-analysis A quantitative method for synthesising the results of a set of studies on a given topic by describing the distribution of effect sizes, and sometimes by analysing differences in effect sizes related to study characteristics

Multiple-choice format items Test questions for which the student selects from among prescribed choices instead of responding with their own ideas

Performance assessments Assessment tasks that require students to carry out a process or produce a product, and associated scoring schemes that require observation and judgment of the quality of that process or product

Rubrics A set of rules to evaluate the quality of a student performance, typically by specifying levels of quality according to a set of criteria for the performance

Speededness The degree to which the speed of an examinee's responses contributes to their score on a test

References

American Educational Research, Association American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: Authors.

ACT Inc. (2005). *Gender fairness using the ACT*. Retrieved May 5, 2008, from: <www.act.org/path/policy/pdf/gender.pdf>.

Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in mathematics and sciences: The gender differences perspective. *Journal of Educational Psychology*, *88*, 365–377.

Breland, H., & Lee, Y.-W. (2007). Investigating uniform and non-uniform gender DIF in computer-based ESL writing assessment. *Applied Measurement in Education*, *20*, 377–403.

Bridgeman, B., & Schmitt, A. (1997). Fairness issues in test development and administration. In W. W. Willingham, & N. S. Cole (Eds.), *Gender and fair assessment.* Mahwah, NJ: Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Coley, R. J. (2001). *Differences in the gender gap: Comparisons across racial/ethnic groups in education and work*. Princeton, NJ: Educational Testing Service.

Cox, P. (2000). Regional and gender differences in mathematics achievement. *Journal of Research in Rural Education*, *16*, 22–29.

Curley, W. E., & Schmitt, A. P. (1993). Revising SAT-Verbal items to eliminate differential item functioning (CB Rep. No. 93-2; ETS RR-93-61). New York: College Entrance Examination Board.

DeFraine, B., Van Damme, J., & Onghena, P. (2007). A longitudinal analysis of gender differences in academic self-concept and language achievement: A multivariate multilevel latent growth approach. *Contemporary Educational Psychology*, *32*, 132–150.

DeLisle, J., Smith, P., & Jules, V. (2005). Which males or females are most at risk and on what? An analysis of gender differentials within the primary school system of Trinidad and Tobago. *Educational Studies*, *31*, 393–418.

DeMars, C. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*, 279–299.

Duffy, J., Gunther, G., & Walters, L. (1997). Gender and mathematical problem solving. *Sex Roles, 37*, 477–494.

Ethington, C. (1990). Gender differences in mathematics: An international perspective. *Journal for Research in Mathematics Education, 21*, 74–80.

Francis, B., & Skelton, C. (2005). *Reassessing gender and achievement: Questioning contemporary key debates*. London: Routledge.

Goldberg, P. (1968). Are women prejudiced against women? *Transaction, 5*, 28–30.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*, 47–77.

Institute of Education Sciences, United States Department of Education. (2007). *National Assessment of Educational Progress at Grades 4 and 8: Reading 2007. NCES Report No. 2007-496*. Retrieved May 1, 2008, from <nces.ed.gov/nationsreportcard/pdf/main2007/2007496.pdf>.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education (Revised)*. Washington, DC: Author.

Klecker, B. M. (2006). The gender gap in NAEP fourth-, eighth-, and twelfth-grade reading scores across years. *Reading Improvement, 43*, 50–56.

Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation, 32*, 317–344.

Nowell, A., & Hedges, L. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles, 39*, 21–43.

Organisation for Econonic Co-operation and Development (OECD). (2004). *Problem Solving for Tomorrow's World—First measures of cross-curricular competencies from PISA 2003*. Paris: Author.

Organisation for Economic Co-operation and Development (OECD). (2007). *PISA 2006: Science competencies for tomorrow's world. Volume 1: Analysis. Paris: Author*.

Robertson, I. (2005). Issues related to curriculum, policy, and gender raised by national and international surveys of achievement in mathematics. *Assessment in Education, 12*, 217–236.

Sammons, P. (1995). Gender, ethnic and socio-economic differences in attainment and progress: A longitudinal analysis of student achievement over 9 years. *British Educational Research Journal, 21*, 465–485.

Swim, J., Borgida, E., Maruyama, G., & Myers, D. G. (1989). Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin, 105*, 409–429.

Wang, N., & Lane, S. (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education, 9*, 175–199.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, NJ: Erlbaum.

Woodfield, R., Earl-Novell, S., & Solomon, L. (2005). Gender and mode of assessment at university: Should we assume female students are better suited to coursework and males to unseen examinations? *Assessment and Evaluation in Higher Education, 30*, 35–50.