The Alberta Journal of Educational Research *Vol. 52, No. 1, Spring 2006, 4-15*

Gregory A. Pope
Carolyn Wentzel
Brigitta Braden

and

Jordan Anderson
Alberta Education

Relationships Between Gender and Alberta Achievement Test Scores During a Four-Year Period

The purpose of this study was to investigate statistical relationships between gender and Alberta Achievement Testing Program scores. Achievement test scores from grades 3, 6, and 9 in all subject areas were investigated during a four-year period. Results showed statistically significant positive correlations between gender and scores in most subject areas. Girls tended to outperform boys in language-based tests (i.e., writing and reading assignments), whereas boys tended to score higher than girls in mathematics and science. Correlation effect sizes were in the small range for the machine-scored components and near the medium size range for the constructed-response components. In general, the results show that small relationships exist between gender and achievement test scores in Alberta.

Cette recherche avait comme objectif d'étudier les rapports statistiques entre le genre et les résultats aux examens de rendement scolaire de l'Alberta. On a étudié les résultats aux examens de la 3e, 6 e et 9 e année pour toutes les matières sur une période de quatre ans. L'analyse révèle des corrélations positives statistiquement significatives entre le genre et les résultats aux examens pour la plupart des matières. Les filles obtenaient de meilleurs résultats que les garçons aux examens portant sur la langue (c.-à-d., les tâches de lecture et de rédaction), alors que les garçons obtenaient de meilleurs résultats aux examens en mathématiques et en science. L'ampleur de l'effet était petite pour les composantes à correction automatique et moyenne pour les questions à réponse construite. En général, les résultats indiquent une dépendance statistique faible entre le genre et le rendement aux examens du Ministère en Alberta.

Research in the area of statistical relationships between gender and student academic performance is abundant, but the findings tend to be inconsistent. In addition, gender typically accounts for small amounts of the variation in test

Greg Pope was a senior psychometrician with the Learner Assessment Branch of Alberta Education. He is currently on leave working as a senior psychometrician for Castle Rock Research Corp.
Carolyn Wentzel is a senior psychometrician with the Learner Assessment Branch.
Brigitta Braden was the Examination Manager for grade 9 humanities with the Learner Assessment Branch and is now a curriculum specialist with Castle Rock Research Corp.
Jordan Anderson was the Examination Manager for grades 6 and 9 and the Integrated Occupational Program (IOP) science with the Learner Assessment Branch and is currently teaching grade 6 and junior high at High Park School in Stony Plain, Alberta.

scores. Willingham, Cole, Lewis, and Leung (1997) examined 74 standardized tests in 15 categories at the grade 12 level and found that gender differences accounted for no more than 1% of the variation in test scores. Gender differences in performance were examined on three United States standardized tests; differences in most subject areas were small and comparable in magnitude and direction across approximately 30 years of testing (Han & Hoover, 1994). Boys tended to score higher in mathematics computation, whereas girls scored somewhat higher in the upper grades on verbal testing. A substantial gender disparity in mathematics and science performance favoring boys was found among 17-year-old students in the US as measured by the National Assessment of Educational Progress (NAEP, Smith, Young, Bae, Choy, & Alsalam, 1997).

The grade 8 results of the 1995 Third International Mathematics and Science Study (TIMSS; Beaton et al., 1996a) showed statistically significant mean gender differences for Canadian students in the sciences; boys demonstrated higher performance than girls. The grade 7 science results for Canadian students showed similar outcomes in that boys scored significantly higher than girls. The 1995 TIMSS study, which focused on mathematics (Beaton et al., 1996b) did not demonstrate statistically significant gender differences for either grade 7 or 8 Canadian students. The 1995 TIMSS Mathematics Achievement in the Primary School Years study (Mullis et al., 1997) examined grades 3 and 4 performance, finding statistically significant mean differences in favor of boys only at the grade 3 level for Canadian students. The 1995 TIMSS Science Achievement in the Primary School Years study (Martin et al., 1997) also examined grades 3 and 4 performance and, parallel to the findings for mathematics, the only significant mean difference was at the grade 3 level in favor of Canadian boys. In the final year of secondary school, significant gender differences were evident among Canadian students for both mathematics and science in favor of boys (Mullis, Martin, Fierros, Goldberg, & Sternler, 2000).

The Program for International Student Assessment (PISA; Bussiere et al., 2001) found significant gender differences among Canadian students in favor of girls on the reading literacy component. Canadian boys scored higher than girls on PISA mathematics literacy; no significant difference for gender was found on PISA science literacy.

The School Achievement Indicators Program (SAIP) involves the pan-Canadian assessment of 13- and 16-year-old students in the areas of mathematics, science, and reading and writing. All subject areas contain items that signify five levels of performance ranging from early stages of competence (Level 1) to full competence (Level 5). These assessments facilitate comparisons to help determine whether students across Canada reach similar levels of performance at about the same age and whether gender differences exist across performance levels (Council of Ministers of Education, Canada [CMEC], 1995). The 1996 Mathematics Assessment (CMEC, 1997) found that girls outperformed boys at the lower levels of competence, but boys outperformed girls at higher competence levels. The 1998 SAIP Reading and Writing Assessments (CMEC, 1999) demonstrated statistically significant gender differences in favor of girls for both 13- and 16-year-old Canadian students in each assessment area across all levels of achievement. The 1999 SAIP for science results indicated few statistically significant gender differences for 13- and 16-year-old Canadian students

on both the written assessment and the practical task portions of the examination (CMEC, 2000). The gender differences found were mainly in the direction of girls outperforming boys, with the sole exception of boys scoring higher than girls at the Level 4 science written assessment.

The TIMSS, PISA, and SAIP assessments offer a Canadian context for gender differences on quality standardized tests with robust sample sizes. These national and international assessments tailor their item coverage to a general curriculum rather than specifically to the Alberta educational system. Nevertheless, the gender results from these studies show patterns with Canadian students that are consistent with studies conducted in the US and elsewhere.

More recently Pope, Wentzel, and Cammaert (2002) investigated gender relationships with regard to performance on grade 12 diploma examinations in Alberta and found that gender had a small statistical and practical effect on diploma examination scores (i.e., less than 1% on average). The authors found larger relationships between gender and teacher-awarded scores.

Historically there has been interest in Alberta in the relationships between gender and student performance on provincial achievement tests at the elementary and junior high school levels. Although the supposition exists that boys outperform girls in areas such as social studies, mathematics, and science, these beliefs have yet to be thoroughly investigated empirically in the context of Alberta.

The purpose of this study was empirically to investigate gender relationships with regard to Alberta achievement test scores in grades 3, 6, and 9 over four years. Specifically, this study endeavored to identify whether relationships exist between gender and scores, the direction of such relationships if any, and the statistical and practical importance of these relationships. The objective of this study is to contribute to the field of educational research by providing information about the previously unexamined role of gender in the context of Alberta standardized test scores at the elementary and junior high school levels.

Method

The province of Alberta maintains an elementary and junior high school-level testing program called the Alberta Achievement Testing Program that comprises a total of 12 criterion-referenced standardized tests administered at the grades 3, 6, and 9 levels. At grade 3, performance in both mathematics and language arts are tested; at grades 6 and 9, mathematics, language arts, social studies, and science abilities are assessed. Three of the 12 examinations are offered in English only (English Language Arts 3, 6, and 9), two are offered in French only (French Language Arts 6 and 9), and seven are offered in English as well as French translations (Mathematics 3, 6, and 9, Science 6 and 9, and Social Studies 6 and 9).

The design of the Alberta Achievement Testing Program is intended to serve three main purposes: (a) to determine if students are learning what they are expected to learn; (b) to report to Albertans how well students have achieved provincial standards at given points in their schooling; and (c) to assist schools, school authorities, and the province in monitoring and improving student learning. Student achievement is reported by means of two assess-

ment standards, the *acceptable standard* and the *standard of excellence*. Approximately half of the items on each achievement test are repeated the following year; statistical equating techniques are then employed to adjust the cut scores of both standards. This process ensures comparability of test results between two consecutive administrations. The tests are administered each year in June.

Achievement tests are composed of one or two sections, a machine-scored section that consists of multiple-choice and numeric-response items (i.e., a type of item where numeric responses are written in cells and then numbered bubbles are shaded), and an open-ended constructed-response section. The English and French language arts courses have both multiple-choice and constructed-response items; all other achievement tests have only multiple-choice or multiple-choice and numeric-response items.

The items for each achievement test are developed by the Learner Assessment Branch and follow the guidelines presented in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999). These standards include instructions for creating items, and consequently tests, that contain gender-neutral language, content, symbols, words, and phrases.

Analyses conducted for this study include data from the English and French Language Arts Achievement Tests for all Albertan students who had valid scores. Analyses of data resulting from Mathematics, Science, and Social Studies Achievement Tests included only students who wrote the English versions of the tests and had valid test scores. In both cases the entire writing population was used to eliminate the possibility of sampling bias. Given the criteria stated above, data from the June 1999, 2000, 2001, and 2002 Achievement Test administrations were evaluated for inclusion in this study.

Correlation coefficients were used to investigate relationships between gender and achievement and to examine the percentage of variation in scores explained by gender. Pearson correlation coefficients were calculated between gender and achievement test scores. Cohen's (1992) effect size guidelines are used throughout this study to describe the magnitude of the correlation coefficients; 0.10, 0.30, and 0.50 represent small, medium, and large effect sizes respectively. Although the magnitude of these effect sizes was designed as a general guideline, it is reasonable to believe that in the context of Alberta test scores, these effect size ranges are appropriate for interpreting the practical importance of the results (i.e., correlations in the small effect size range are of little practical importance because the percent of variation accounted for is very small). In addition to the Alberta context, Cohen's effect size ranges are consistent with the interpretation of relationships between gender and test scores found in the research literature (Willingham & Cole, 1997a).

To elaborate on the practical interpretation of small effect sizes, one would expect factors such as handedness (left or right) to account for small amounts of variation in scores (thus having small effect sizes), as variables such as this should have no logical interaction with students' performance on standardized tests. For example, if a study found that 50% of the variation in students' standardized test scores was accounted for by handedness, a teacher could simply group the students according to use of right hand or left hand to

determine approximately their future performance on the test. In fact we would expect student handedness to account for virtually no variability in standardized test scores.

Females were coded as *1* and boys were coded as *2*; therefore, a positive correlation indicates a relationship in the direction of boys, a negative correlation is a relationship in favor of girls. For example, a statistically significant correlation of 0.11 between gender and Achievement Test scores indicates that boys scored higher than girls (i.e., a statistically significant but small mean difference between boys and girls in the direction of boys is present). Means were not presented in this study due to the redundancy of such reporting (i.e., reporting correlation coefficients provides us with parallel overall information). *N*-ratios were calculated using the ratio of number of female students to number of male students writing each test.

Results

Tables 1 through 4 display the Pearson correlation coefficients for each course, by year.

The results for all Achievement Tests administered in 2002 are shown in Table 1. All tests, with the exception of Social Studies 6, demonstrated statistically significant results with girls scoring higher than boys in English Language Arts 3, 6, and 9 and French Language Arts 6 and 9. Boys scored higher than girls in all other courses. The percentage of score variation accounted for by gender ranged from 0.05% in Mathematics 9 to 5.20% for the constructed-response section of English Language Arts 9. These results were in the small effect size range.

Table 2 presents the results for all Achievement Tests during the 2001 administration. All tests were statistically significant with the exception of Social Studies 6 and Mathematics 9. Girls outperformed boys in English Language Arts 3, 6, and 9 and French Language Arts 6 and 9, whereas boys scored higher than girls in other subjects. The percentage of variation in scores explained by gender ranged from 0.11% in Social Studies 9 to 6.30% for the constructed-response section of French Language Arts 9. Small effect sizes were observed for all correlations.

The results for all Achievement Tests administered in 2000 are shown in Table 3. All participants demonstrated statistically significant results, with girls scoring higher than boys in the language arts subjects and boys scoring higher than girls in all other subjects. The percentage of score variation explained by gender ranged from 0.06% in Mathematics 6 to 6.00% for the constructed-response section of English Language Arts 9. All significant correlations were in the small effect size range.

Table 4 presents the results for all Achievement Tests during the 1999 administration. Findings were similar to those observed from the 2000 administration; all participants showed statistically significant results, with girls outperforming boys in the language arts courses and boys outperforming girls in all other subjects. The percentage of variation in scores explained by gender ranged from 0.01% in Social Studies 6 to 5.29% in the constructed-response section of English Language Arts 9. Small effect sizes were observed for all correlations.

Table 1
Correlations Between Gender and Alberta Achievement Test Scores, June 2002 Session

Achievement Courses	*Total N*	*N ratio*	*Gender by achievement test score*		
			r	*Significance*	*Percent of variation in scores explained by gender*
English Language Arts 3	38,370	0.969	–0.123	$p<0.001$	1.51%
Multiple-choice	38,370	0.969	–0.069	$p<0.001$	0.48%
Constructed-response	38,370	0.969	–0.158	$p<0.001$	2.50%
Mathematics 3	36,197	0.945	0.070	$p<0.001$	0.49%
English Language Arts 6	41,114	0.973	–0.151	$p<0.001$	2.28%
Multiple-choice	41,114	0.973	–0.070	$p<0.001$	0.49%
Constructed-response	41,114	0.973	–0.213	$p<0.001$	4.54%
Mathematics 6	39,210	0.958	0.051	$p<0.001$	0.26%
Science 6	39,054	0.963	0.081	$p<0.001$	0.66%
Social Studies 6	38,843	0.964	–0.002	$p=0.673$	–
French Language Arts 6	2,188	1.17	–0.144	$p<0.001$	2.07%
Multiple-choice	2,188	1.17	–0.059	$p=0.006$	0.35%
Constructed-response	2,188	1.17	–0.211	$p<0.001$	4.45%
English Language Arts 9	38,001	0.965	–0.185	$p<0.001$	3.42%
Multiple-choice	38,001	0.965	–0.102	$p<0.001$	1.04%
Constructed-response	38,001	0.965	–0.228	$p<0.001$	5.20%
Mathematics 9	35,891	0.950	0.022	$p<0.001$	0.05%
Science 9	36,444	0.947	0.100	$p<0.001$	1.00%
Social Studies 9	36,278	0.946	0.036	$p<0.001$	0.13%
French Language Arts 9	2,120	1.24	–0.098	$p<0.001$	0.96%
Multiple-choice	2,120	1.24	–0.019	$p=0.392$	–
Constructed-response	2,120	1.24	–0.170	$p<0.001$	2.89%

Note. *N* ratio=ratio of the number of female students to male students.
Significance=2-tailed.
Gender coding female=1, male=2.

Discussion

The results presented in this article demonstrate that in general, gender has a very small statistical affect on achievement test scores. This finding is consistent with the current literature on standardized test score gender differences at the lower grade levels (Willingham et al., 1997). Relationships between gender and test scores tend to be very small for most courses that contain only machine-scored test items.

The language arts courses, which include both constructed-response and machine-scored test items, were consistently found to have the strongest relationships between gender and test scores. The constructed-response section for these tests was the major contributing factor in the percentage of variation accounted for by gender in the total test score, with girls always outperforming boys on the writing component. The amount of variation in constructed-re-

Table 2
Correlations Between Gender and Alberta Achievement Test Scores, June 2001 Session

Achievement Courses	*Total N*	*N ratio*	*Gender by achievement test score*		
			r	*Significance*	*Percent of variation in scores explained by gender*
English Language Arts 3	40,170	0.972	–0.104	$p<0.001$	1.08%
Multiple-choice	40,170	0.972	–0.050	$p<0.001$	0.25%
Constructed-response	40,170	0.972	–0.146	$p<0.001$	2.13%
Mathematics 3	37,803	0.957	0.073	$p<0.001$	0.53%
English Language Arts 6	41,185	0.955	–0.160	$p<0.001$	2.56%
Multiple-choice	41,185	0.955	–0.083	$p<0.001$	0.69%
Constructed-response	41,185	0.955	–0.213	$p<0.001$	4.54%
Mathematics 6	39,324	0.941	0.042	$p<0.001$	0.18%
Science 6	39,324	0.942	0.093	$p<0.001$	0.86%
Social Studies 6	39,211	0.944	0.005	$p=0.352$	–
French Language Arts 6	2,212	1.19	–0.139	$p<0.001$	1.93%
Multiple-choice	2,212	1.19	–0.045	$p=0.033$	0.20%
Constructed-response	2,212	1.19	–0.218	$p<0.001$	4.75%
English Language Arts 9	38,767	0.985	–0.185	$p<0.001$	3.42%
Multiple-choice	38,767	0.985	–0.108	$p<0.001$	1.17%
Constructed-response	38,767	0.985	–0.230	$p<0.001$	5.29%
Mathematics 9	36,777	0.965	0.009	$p=0.090$	–
Science 9	37,275	0.961	0.070	$p<0.001$	0.49%
Social Studies 9	37,014	0.961	0.033	$p<0.001$	0.11%
French Language Arts 9	2,119	1.39	–0.197	$p<0.001$	3.88%
Multiple-choice	2,119	1.39	–0.107	$p<0.001$	1.14%
Constructed-response	2,119	1.39	–0.251	$p<0.001$	6.30%

Note. *N* ratio=ratio of the number of female students to male students.
Significance=2-tailed.
Gender coding female=1, male=2.

sponse scores explained by gender was consistently (i.e., on every test that contained a constructed-response section) much greater than that accounted for by gender in multiple-choice scores. For example, the amount of variability due to gender was 9.3 times higher for constructed-response scores compared with multiple-choice scores on the June 2002 English Language Arts 6 test.

The Alberta achievement tests containing constructed-response components are intended to assess writing skills; as such, the gender relationships found in this study are in accord with earlier research, which indicates that in general girls outperform boys on measures of writing skills (CMEC, 1999; Willingham & Cole, 1997a). It is unlikely that the relationships between gender and constructed-response scores are the result of bias in question content as these relationships have been observed consistently over the four years of testing investigated in this study. In addition, girls scored higher than boys on

Table 3
Correlations Between Gender and Alberta Achievement Test Scores, June 2000 Session

Achievement Courses	*Total N*	*N ratio*	*Gender by achievement test score*		
			r	*Significance*	*Percent of variation in scores explained by gender*
English Language Arts 3	40,428	0.98	–0.101	$p<0.001$	1.02%
Multiple-choice	40,428	0.98	–0.053	$p<0.001$	0.28%
Constructed-response	40,428	0.98	–0.134	$p<0.001$	1.80%
Mathematics 3	37,978	0.96	0.049	$p<0.001$	0.24%
English Language Arts 6	40,891	0.98	–0.169	$p<0.001$	2.86%
Multiple-choice	40,891	0.98	–0.103	$p<0.001$	1.06%
Constructed-response	40,891	0.98	–0.208	$p<0.001$	4.33%
Mathematics 6	38,861	0.96	0.024	$p<0.001$	0.06%
Science 6	38,483	0.96	0.067	$p<0.001$	0.45%
Social Studies 6	38,722	0.96	0.014	$p=0.005$	0.02%
French Language Arts 6	2,339	1.32	–0.143	$p<0.001$	2.04%
Multiple-choice	2,339	1.32	–0.042	$p=0.041$	0.18%
Constructed-response	2,339	1.32	–0.238	$p<0.001$	5.66%
English Language Arts 9	39,418	0.98	–0.163	$p<0.001$	2.66%
Multiple-choice	39,418	0.98	–0.065	$p<0.001$	0.42%
Constructed-response	39,418	0.98	–0.245	$p<0.001$	6.00%
Mathematics 9	37,419	0.96	0.036	$p<0.001$	0.13%
Science 9	37,687	0.96	0.105	$p<0.001$	1.10%
Social Studies 9	37,588	0.95	0.028	$p<0.001$	0.08%
French Language Arts 9	2,212	1.34	–0.138	$p<0.001$	1.90%
Multiple-choice	2,212	1.34	–0.074	$p<0.001$	0.55%
Constructed-response	2,212	1.34	–0.178	$p<0.001$	3.17%

Note. *N* ratio=ratio of the number of female students to male students.
Significance=2-tailed.
Gender coding female=1, male=2.

the machine-scored component of these language arts tests. This finding suggests that girls may excel in the language content areas of courses compared with boys, although the gender effects are relatively small.

The association between girls and higher scores in the language arts areas has been explained in the literature as gender differences in innate skills. Earlier research has shown that girls are more verbally apt than boys throughout elementary grades (Halpern, 1992). Maccoby and Jacklin (1974) stated that girls tend to show a stronger aptitude toward verbal skills than boys at approximately age 10 and that this aptitude continues through their high school and college years. Cleary (1992) found that girls outperformed boys on verbal tests from early to later grades. Evidence also suggests that girls tend to report greater intentions to pursue college study in the areas of language and humanities (Dwyer & Johnson, 1997). Although suggestions of a causal as-

Table 4
Correlations Between Gender and Alberta Achievement Test Scores, June 1999 Session

Achievement Courses	*Total N*	*N ratio*	*Gender by achievement test score*		
			r	*Significance*	*Percent of variation in scores explained by gender*
English Language Arts 3	41,290	0.96	–0.109	$p<0.001$	1.19%
Multiple-choice	41,290	0.96	–0.060	$p<0.001$	0.36%
Constructed-response	41,290	0.96	–0.142	$p<0.001$	2.02%
Mathematics 3	38,855	0.95	0.019	$p<0.001$	0.04%
English Language Arts 6	40,068	0.96	–0.187	$p<0.001$	3.50%
Multiple-choice	40,068	0.96	–0.115	$p<0.001$	1.32%
Constructed-response	40,068	0.96	–0.230	$p<0.001$	5.29%
Mathematics 6	37,765	0.95	0.032	$p<0.001$	0.10%
Science 6	37,836	0.95	0.073	$p<0.001$	0.53%
Social Studies 6	37,725	0.95	0.012	$p=0.017$	0.01%
French Language Arts 6	2,322	1.14	–0.136	$p<0.001$	1.85%
Multiple-choice	2,322	1.14	–0.068	$p=0.001$	0.46%
Constructed-response	2,322	1.14	–0.182	$p<0.001$	3.31%
English Language Arts 9	39,026	0.97	–0.162	$p<0.001$	2.62%
Multiple-choice	39,026	0.97	–0.091	$p<0.001$	0.83%
Constructed-response	39,026	0.97	–0.213	$p<0.001$	4.54%
Mathematics 9	36,652	0.95	0.028	$p<0.001$	0.08%
Science 9	37,102	0.94	0.097	$p<0.001$	0.94%
Social Studies 9	37,089	0.95	0.033	$p<0.001$	0.11%
French Language Arts 9	2,331	1.41	–0.197	$p<0.001$	3.88%
Multiple-choice	2,331	1.41	–0.151	$p<0.001$	2.28%
Constructed-response	2,331	1.41	–0.203	$p<0.001$	4.12%

Note. *N* ratio=ratio of the number of female students to male students.
Significance=2-tailed.
Gender coding female=1, male=2.

sociation would be premature, clearly innate verbal ability and selective subject areas of interest appear to play a role in girls outperforming boys on verbal assessments. Further research to investigate the relationships between innate gender factors and performance on Alberta assessments would help to bridge this knowledge gap.

This study demonstrated that with few exceptions, boys outperformed girls by a small margin in the mathematics, science, and social studies courses in all grades. Girls consistently scored higher than boys on all language arts tests. This pattern was observed across all four years of achievement test data included in this study. Earlier research has suggested that in general, boys tend to outperform girls on standardized tests in the areas of mathematics and sciences (Beaton et al., 1996a, 1996b; Mullis, Owen, & Phillips, 1990). Research also indicates that this small gender difference in the math/sciences is relative-

ly constant from grades 4 to 12 (Willingham & Cole, 1997a). Pope et al. (2002) found small gender differences in favor of boys in several of the math/science subjects on Alberta grade 12 standardized tests. These relationships could be partly explained by male preferences for mathematics and sciences, as boys tend to report greater intentions to pursue college study in the mathematics, physical sciences, and computer sciences (Dwyer & Johnson, 1997). In the current study, the relationships between gender and test scores in mathematics and science courses favoring boys are far less pronounced than the significant relationships between gender and constructed-response scores favoring girls.

Consistently across the four years of test administration, the most negligible relationship between gender and test scores was found at the grade 6 level in social studies. At the grade 9 level the relationships were also very small. Some research in the area of gender differences in the social sciences has shown results similar to those of the current study with boys scoring slightly higher on tests than girls (Willingham et al., 1997). However, in the context of Alberta, Pope et al. (2002) found relationships between gender and social studies diploma examination scores at the grade 12 level that were greater in magnitude than any other subjects at that level (e.g., mathematics, sciences, and language arts). Examination of school-awarded scores at the same grade level indicated an opposing pattern. The relationship between gender and school-awarded scores for social studies was negligible, but relationships of relatively larger magnitude were found in the other subject areas. These findings underscore the fact that a clear depiction of the interaction between gender and assessment scores at various grade levels does not exist. Future exploration in this area in the context of Alberta would be invaluable toward a better understanding of gender difference, if apparent, and informing future teaching practices with regard to gender and course content.

The findings of this study provide useful information for stakeholders of the Alberta achievement testing program (e.g., students, parents, teachers, and school and jurisdiction administrators). A question commonly asked by members of these groups is "Do some subjects, as assessed by Alberta achievement tests, contain inherent bias towards males or females?" Specific subject areas such as mathematics and social studies have reputations as courses that appeal more to boys than girls. Before the current study, detailed empirical analyses had not been conducted to assess potential bias in Alberta achievement tests. The present study should aid toward assuring stakeholders that gender plays a very small role in student performance on Alberta achievement tests. The strongest relationships found between gender and test scores were in the constructed-response sections of the language arts tests, with girls outperforming boys. This finding warrants future research in order to provide instructional direction to teachers toward assisting boys to develop stronger writing skills.

Limitations of this study include the lack of psychological information to augment the gender findings. Specifically, it would be useful to investigate in more detail the psychological mitigating factors as to why small relationships were found between gender and Alberta achievement test scores. A controlled study where personality tests and surveys are administered to students may help to flesh out the psychological underpinnings. Another limitation of this study is the lack of teacher-awarded assessments of the students' performance.

Pope et al. (2002) examined relationships between gender and the Alberta grade 12 diploma examination scores as well as teacher-awarded scores (school-awarded scores). We could not readily use this additional information for achievement tests because teacher-awarded scores are not collected systematically by the Alberta government. A future empirical study in which teacher assessments of student performance are collected and analyzed, in combination with student achievement test scores and gender information, may yield interesting findings that can be contrasted with the grade 12 findings reported by Pope et al.

The current study may serve as a basis for future research in that the results provide a foundation for a proposed three-stage approach to exploratory investigations of gender differences. The first stage would involve an empirical study similar to the current research, where differences between gender are investigated at a total test score level. Should the findings yield results in the medium or large effect size ranges, a second stage would be undertaken to identify individual items that display significant levels of gender DIF (differential item functioning). The third stage would consist of a detailed content analysis of the items flagged as displaying DIF in order to determine whether the items are biased toward one gender group or whether the DIF is a result of impact (i.e., natural gender differences). If stage one indicates a very small relationship between gender and test scores, this finding provides evidence that gender bias is not an issue. Consequently, few items or none would be expected to display DIF. Use of this multistep approach to investigate the role of gender on standardized tests can reduce the need for researchers to allocate resources and time in conducting a detailed DIF study where probably nothing of practical importance would be found.

The results presented here demonstrate that very small relationships exist between gender and achievement test scores in Alberta. Girls tend to have an advantage over boys in regard to language-based tests (i.e., writing and reading assignments), whereas boys tend to score higher than girls in the areas mathematics and science. The results of this study contribute to the field of education by providing information on the role of gender in the context of Alberta standardized test scores at the elementary and junior high school levels.

References

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

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E, J., Smith, T, A., & Kelly, D.L. (1996a). *Science achievement in the middle school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1996b). *Mathematics achievement in the middle school years: IEA's third international mathematics and science study (TIMSS)*. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Bussiere, P., Cartwright, F., Crocker, R., Ma, X., Oderkirk, J., & Zhang, Y. (2001). *Measuring up: The performance of Canada's youth in reading, mathematics and science*. Ottawa: Human Resources Development Canada, Council of Ministers of Education Canada, and Statistics Canada.

Cleary, T.A. (1992). Gender differences in aptitude and achievement test scores. *Sex equity in educational opportunity, achievement, and testing: Proceedings of the 1991 ETS invitational conference* (pp. 51-90). Princeton NJ: Educational Testing Services.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Council of Ministers of Education, Canada (CMEC). (1995). *About SAIP* [On-line]. Toronto, ON: Author. Available: http://www.cmec.ca/saip/about.htm

Council of Ministers of Education, Canada (CMEC). (1997). *Report on mathematics assessment school achievement indicators program (SAIP) 1997*. Toronto, ON: Author.

Council of Ministers of Education, Canada (CMEC). (1999). *Report on reading and writing assessment school achievement indicators program (SAIP) 1998*. Toronto, ON: Author.

Council of Ministers of Education, Canada (CMEC). (2000). *Report on science assessment school achievement indicators program (SAIP) 1999*. Toronto, ON: Author.

Dwyer, C.A., & Johnson, L.M. (1997). Grades, accomplishments, and correlates. In W. Willingham & N.S. Cole (Eds.), *Gender and fair assessment* (pp. 127-156). Mahwah, NJ: Erlbaum.

Halpern, D.F. (1992). *Sex differences in cognitive abilities* (2nd ed.). Hillsdale, NJ: Erlbaum.

Han, L., & Hoover, H.D. (1994, April). *Gender differences in achievement test scores*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

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

Martin, M.O., Mullis, I.V.S., Beaton, A.E., Gonzalez, E.J., Smith, T.A., & Kelly, D.L. (1997). *Science achievement in the primary school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Mullis, I.V.S., Martin, M.O., Beaton, A.E., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1997). *Mathematics achievement in the primary school years: IEA's third international mathematics and science study (TIMSS).* Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

Mullis, I., Martin, M., Fierros, E., Goldberg, A., & Sternler, S. (2000). *IEA's third international mathematics and science study (TIMSS). Gender differences in achievement*. Boston, MA: TIMSS International Study Center, Boston College.

Mullis, I.V., Owen, E.H., & Phillips, G.W. (1990). *America's challenge: Accelerating academic achievement—A summary of findings from 20 years of NAEP*. Princeton, NJ: Educational Testing Service, National Assessment of Educational Progress.

Pope, G.A., Wentzel, C., & Cammaert, R. (2002). Relationships between gender and Alberta diploma scores. *Alberta Journal of Educational Research, 48*, 275-286.

Smith, T.M., Young, B.A., Bae, Y., Choy, S.P., & Alsalam, N. (1997). *The condition of education 1997. Research Report No. NCES 97-388*. Washington, DC: US Department of Education, National Center for Education Statistics.

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

Willingham, W.W., & Cole, N.S. (1997b). Fairness issues in test design and use. In W. Willingham & N.S. Cole (Eds.), *Gender and fair assessment* (pp. 227-345). Mahwah, NJ: Erlbaum.

Willingham, W.W., Cole, N.S., Lewis, C., & Leung, S.W. (1997). Test performance. In W. Willingham & N.S. Cole (Eds.), *Gender and fair assessment* (pp. 55-126). Mahwah, NJ: Erlbaum.