*Caribbean Curriculum*
*Vol. 11, 2004, 23–55*

# RECONSIDERING THE CONSEQUENCES
## Gender Differentials in Performance and Placement in the 2001 SEA

*Jerome De Lisle* and *Peter Smith*

This paper provides an analysis of the gender fairness and consequences associated with the test design used for the 2001 Secondary Entrance Assessment (SEA) in Trinidad and Tobago. It is argued that the rationale for choosing the SEA test design emphasized the usefulness and purpose of the selection instrument, but failed to consider one significant consequence: the likelihood of adverse impact resulting from large performance differentials in favour of females. The study also tests the hypotheses that gender differences are (1) institution-specific and (2) vary across ability groups. The major findings were that patterns of gender inequity were complex and sometimes even contradictory, with females favoured on SEA composite total score, language arts, and creative writing and males favoured on the placement process. However, males and females performed similarly in mathematics. An analysis across different ability groups indicated that large differentials favouring females were more likely among students below the 50th percentile. On the other hand, among higher achievers, males performed just as well as females. The gender fairness of five alternative SEA test designs was evaluated using Willingham's (1999) social matrix.

*The point is that the functional worth of the testing depends not only on the degree to which the intended purposes are served but also the consequences of the outcomes produced, because the values captured in the outcomes are at least as important as the values unleashed in the goals.* (Messick, 1989, p. 85)

## Introduction

### Gender differentials in favour of females: A growing problem in the English-speaking Caribbean

Large gender differentials favouring females are a major concern at all levels of the education system within a number of Caribbean territories, including Jamaica, Dominica, Barbados, and Trinidad and Tobago (Bailey, 2000; Goldberg & Bruno, 1999; Kutnick, Jules, & Layne, 1997; Layne & Kutnick, 2001). However, this phenomenon has also been observed in First World countries such as the UK, Australia, and New Zealand (Alloway & Gilbert, 1997; Gallagher, 1997; Harker, 2000). In the research to date, the primary focus has been on achievement and access as indicators of educational performance, with differences in enrolment ratios and tests scores most often emphasized (Leo-Rhynie, 1999).

In the English-speaking Caribbean and elsewhere, some have approached the issue from the singular perspective of male underachievement, arguing that males are at a disadvantage in schooling (Goldberg & Bruno, 1999; Miller, 1991). However, it might well be that patterns of gender inequity are complex and contradictory, varying for each sub-group across attainment levels and contexts (De Lisle & Pitt-Miller, 2002; Elwood, 1999b; Gorard, Rees, & Salisbury, 1999; Warrington & Younger, 2000). Nevertheless, fairness and equity are critical concepts in the search for social justice, and therefore it is important to determine which factors contribute to differential performance (Gipps & Murphy, 1994). Indeed, the opportunity to develop the full potential of each gender is central to efforts at human resource development (Behrman, 1996). As such, one might argue that analyzing the nature of gender inequity within the education system is a prerequisite to planning a 2020 vision for Trinidad and Tobago.

### Reconsidering the research: Issues of measurement and focus

Despite the increasing number of empirical studies analyzing gender differences in the English-speaking Caribbean, two notable weaknesses are readily apparent. The first is the failure to report statistical measures that allow an evaluation of the practical significance of differentials. Although this is the most critical aspect of the analysis, in the past, some studies have made judgements based solely upon the raw difference

between means (Rampersad, 1999). On the other hand, although newer studies frequently report tests of statistical significance, effect size measures are rarely included (Bailey, 2000; Layne & Kutnick, 2001). Practical significance, however, can only be assessed through a measure of effect size (Daniel, 1988; De Lisle & Pitt-Miller, 2002; Fan, 2001). Indeed, it is critical for researchers to distinguish between statistical and practical significance, especially in studies with large sample sizes, because statistically significant differences will be obtained even when the magnitude of differentials are, in fact, relatively small (Thompson, 2002). For this reason, many publication manuals now make the reporting of effect size mandatory (American Psychological Association [APA], 2001; Daniel, 1988).

The most common and useful effect size measure in the study of gender differentials is the standard mean difference, Cohen's d (Pomplun & Sundbye, 1999). This index is obtained by subtracting the mean score of females from the mean score of males and dividing it by the pooled standard deviation (Willingham & Cole, 1997). Cohen (1988) also provided a metric for interpreting the standard mean difference in terms of practical significance. According to this estimation, d-values of ± 0.2 are considered negligible, 0.2 to 0.5 small, 0.5 to 0.8 medium-sized, and values greater than 0.8 large.

A second notable weakness in the empirical research to date is the limited range of indices used to describe patterns in differential performance. As a result, it is possible that some issues have not been readily apparent and the overall complexity of patterns obscured. To illustrate, Feingold (1992) noted the tendency for the scores of males and females on critical intellectual abilities to differ both in magnitude and variability. He showed that, for any given value of Cohen's d, the gender with the wider distribution of scores would be over-represented at the tails. Thus, differences in score variability will result in differential impact at the high- and low-ability groups. It is always useful, then, to report both a measure of effect size and index of variability when analyzing differential performance by gender. One useful index of variability is the standard deviation ratio (SDR), defined as the ratio of the standard deviation for the female sub-population to the male sub-population (Willingham & Cole, 1997; Witt, Dunbar, & Hoover, 1994). Another useful index for measuring the impact of gender differentials, especially across different ability groups, is the female to male ratio

(F/M ratio). Willingham, Cole, Lewis, and Leung (1997) used the F/M ratio for failing students at different percentiles as a measure of the relative impact of differential performance across the ability range.

While a number of studies have focused upon differential performances at the secondary level, there is a dearth of empirical studies at the all-important primary-secondary transition point. The absence of empirical research at this transition point is alarming because of the high-stakes nature of the decisions. Indeed, it is common in the English-speaking Caribbean to administer a one-shot public examination at this transition point. This choice of selection instrument remains a colonial legacy, which continues despite the growing assessment literacy of educators and the public (Payne & Barker, 1986). Inferences from this one-shot examination are used to make critical life decisions, often resulting in vastly different opportunities and outcomes. For example, in Trinidad and Tobago, placement outcomes might range from a high-achieving seven-year secondary school to a special school for low achievers. By no means can these different outcomes be considered equal, neither can fairness nor meritocracy be automatically assumed, especially without evidence (London, 1989).

With universal secondary education implemented in 2001, it might have appeared that the stakes associated with the placement process were significantly reduced. However, in reality, the prospect of outright failure was replaced by the threat of assignment to a classroom or school for special children. Arguably, this placement decision might be considered somewhat dubious since it is neither standard nor criterion-referenced, but based upon a norm-referenced cut score of 30%. In short, poor performance at 11+ will greatly alter the child's future life chances. It becomes critical, then, to examine the consequences associated with the current 11+ selection procedure. Indeed, the negative impact of assessment was one of the major concerns that led to the change from the Common Entrance Examination (CEE) to the Secondary Entrance Assessment (SEA) in 2001. The negative impact of the CEE was believed to be individual, pedagogical, and curricular, ranging from curriculum distortion to undue anxiety among students. By contrast, it was believed that the SEA would prove a more useful design, with the likelihood of positive impact upon teaching and learning in the primary school. However, in this paper, we will focus upon the adverse consequences that result from use of the SEA. More specifically, we intend to explore

the possibility of differential performance among males and females and its impact upon placement. We argue that this aspect of validity must always be considered when evaluating high-stakes tests in Caribbean societies because of the likelihood of large gender differentials favouring females and the possibility that gender might interact with other variables in predicting performance. We strongly support the argument that if any high-stakes test is to be administered at 11+, it requires a design that is explicitly gender-fair (Chilisa, 2000).

**Consequences and fairness as aspects of validity**

Gender-fairness is a subset of the broader issue of fairness in test design and use. Willingham and Cole (1997) highlighted the importance of validity as a conceptual framework for analyzing test fairness, noting that:

> Validity is an all-encompassing technical standard for judging the quality of the assessment process. Validity includes, for example, the accuracy with which a test measures what it purports to measure, how well it serves its intended function, other consequences of test use, and comparability of the assessment process for different examinees. We see fairness reflected in various aspects of comparable measurement, and anything that reduces fairness tends also to reduce validity. (p. 228)

This broad approach to validity incorporates Messick's (1989) notion of the consequential aspect of testing. Messick (1989, 1994) stressed the need to focus upon the social consequences of test use and argued that both intended and unintended consequences should be considered. Indeed, both Moss (1997) and Broadfoot (2002) have reminded us that, like an emerging Frankenstein's monster, unintended consequences can sometimes outweigh the positive impact of assessment change. In analyzing tests used for selection, Messick (1989) has argued that the entire process should be considered along with alternatives and values that provide the foundation for the test design. This is a useful approach because it suggests that, in examining the consequences of the SEA, both test design and placement should be considered. Indeed, Jules (1994) has provided evidence that a number of ancillary factors are likely predictors of the CEE placement process.

Willingham (1999) developed a social matrix for evaluating fairness in test design and use, consisting of three distinct criteria: usefulness, fairness, and practicality. Usefulness is the intended function or purpose of the tests, and practicality refers to the constraints that impact upon acceptability and feasibility in use. Fairness, the focus of this study, is the differential performances of sub-groups resulting from factors either related or unrelated to the test construct. Willingham argued that in the design of a high-stakes test, the three criteria must be balanced against each another. Earlier, Cole and Moss (1989) developed an alternative framework for evaluating fairness. They used the concept of test bias, defined as the "differential validity of a given interpretation of a test score for any definable, relevant subgroup of test takers" (p. 205). The five categories considered were: (1) constructs in context, (2) content and format, (3) administration and scoring, (4) internal test structure, and (5) external test relationships. Therefore, both frameworks consider construct and format as critical aspects of test design. We believe that in making choices about construct and format, the possibility of adverse impact resulting from differential performance by gender must be given equal weight to any expectation of positive impact.

**Predicting consequences: Factors influencing differential outcomes**

Gender differentials in achievement are influenced by a wide variety of individual, sub-group, and systemic variables. However, the majority of studies in the English-speaking Caribbean highlight primarily sub-group variables, with some theorists perhaps overemphasizing the role of the victim in the resolution of inequity (Figueroa, 2002; Parry, 2000). Nevertheless, there is now a growing awareness of the impact of systemic variables, including institutional characteristics such as school ethos; organizational variables like tracking; and classroom factors like pedagogical style (Evans, 2001; Fuller, Hua, & Snyder, 1994; Kutnick, Jules, & Layne, 1997). A UK study by Daniels, Hey, Leonard, Fielding, and Smith (1999) provides an interesting perspective on the possible role of systemic variables in the creation of gender differentials. This study included schools in which either males or females performed better on English. It was found that some institutions were able to minimize the size and direction of gender differentials by establishing a pedagogic focus on learning rather than on learners, and by fostering collaborative

and supportive environments. On the other hand, the demand of performance pedagogies within competitive classrooms often provoked defensive/subversive attitudes in boys. The possibility that some schools might reduce the differences between male and female performance is a hypothesis worth testing, because it implies that whole-school strategies can be developed to ameliorate the underachievement of specific sub-groups (Younger, Warrington, & Williams, 1999).

One of the more important systemic variables, with a consistent impact across different contexts, is student assessment. Assessment-related factors that may influence gender differentials include construct, task, format, and the stakes involved in testing (Chilisa 2000; Willingham & Cole 1997). An emerging explanatory model suggests that different assessment factors interact with both the learning environment and gendered preferences for working, knowing, and communicating in creating differentials in performance across assessment purposes, formats, and tasks (Elwood, 1999b). For example, one possibility is that females prefer social aspects of learning and may therefore have an advantage in collaborative learning environments and on assessment tasks that emphasize communication and teamwork. On the other hand, it could be that the failure to acknowledge and remediate these gendered preferences accentuates differential performance. For example, some classrooms might inhibit the achievement of males by failing to provide the scaffolding necessary to reach minimal competence on collaborative or performance tasks.

Choice of construct is one of the more important decisions that determine the size and direction of gender differentials. For example, regardless of assessment format, females consistently do better on language-related tasks such as writing and reading, although this advantage is much reduced at both secondary and tertiary level. On the other hand, in mathematics, females often do better on tasks related to computation and knowledge of concepts but are worse at problem solving (Garner & Engelhard, 1999). However, the pattern is less clear in the social and natural sciences (Rampersad, 1999). In many cultures, males have an advantage in the natural sciences, a difference which increases with age (Herbert & George, 1996). Males may also have an advantage in some social sciences, such as geography, where the difference is often sizable and possibly related to differential ability on spatial tasks. Although these differences are often apparent at an early

age, the size of the differentials changes with class level, independent of the assessment used (Gray & Sharp, 2001; Witt, Dunbar, & Hoover, 1994). For example, in writing and language use, females increase their substantial advantage over males up to Form 1, but then the gap remains steady. On the other hand, the gap in favour of males appears to widen in mathematics, science, and geography at secondary school, although the rate of increase is comparatively smaller (Willingham et al., 1997). Kutnick's (1999) study of schools across Barbados suggests that this pattern is also evident in the Caribbean. Specifically, he found that large differentials favouring females in Standard 1 of the primary school were reduced or reversed by Form 2 in secondary school across a range of subjects including science, English, social studies, and mathematics.

There is a substantial and conclusive body of literature that describes the influence of assessment format on both the size and direction of gender differentials. For example, the early work of Murphy (1982) showed that females had a significant advantage on constructed response (CR) items in the British General Certificate of Education (GCE). A similar pattern was also observed in the US Advanced Placement (AP) examinations, although the magnitude of the effect varied by discipline (Breland, Danos, Kahn, Kubota, & Bonner, 1994; Bridgeman & Lewis, 1994). At the same time, differentials in selected response (SR) formats such as multiple choice (MC) tests are often negligible or favour males (Mullins & Greene, 1994). While some have taken these results to mean that MC tests are biased against females, the reverse argument might well hold for CR tests against males. The truth is that no assessment, however authentic, is implicitly gender neutral and the authenticity of an assessment task will not compensate for adverse impact due to large gender differentials favouring one sub-group (Chilisa, 2001; Elwood, 1999b; Gipps & Murphy, 1994).

Gender-influenced format effects have been observed at all levels and on a variety of subjects in basic schooling, including mathematics (Garner & Engelhard, 1999), reading (Pomplun & Sundbye, 1999), and science (DeMars, 1998). Format effects are partly responsible for the female advantage in coursework; however, test stakes and motivation may also have a role in determining differences in performance when scores on coursework and final examinations are compared (DeMars, 2000; Elwood, 1999a). Format effects may be due to construct-related, task-embedded factors such as language skill or verbal fluency in writing, or

to construct-irrelevant factors such as test-wiseness, neatness, handwriting, scoring, and differential reliability (Pomplun & Sundbye, 1999).

These findings have important implications for test design. Moreover, although test design is the first step in the assessment cycle, followed by development, administration, and use, fairness issues permeate all four steps. Nevertheless, decisions at the test design stage often prove critical because this step is intimately interwoven with the others. For example, test development depends upon the table of specifications constructed in test design. The development of items therefore requires frequent cycling between these two stages. Likewise, score reporting and intended test use must be congruent with the design chosen (Willingham & Cole, 1997). In view of the possible impact on the size and direction of gender differentials, the choice of construct and format are major considerations in the design of a fair test (Chilisa, 2000). For example, if an examination samples heavily from constructs and tasks in the verbal-linguistic sphere, large gender differentials favouring females are likely. Similarly, if an examination makes exclusive use of constructed response, gender differentials will be shifted towards females.

> *When the use of a test results in outcomes that affect the life chances or educational opportunities of examinees, evidence of mean test score differences between relevant subgroups of examinees, should, where feasible, be examined for subgroups for which credible research reports mean differences for similar tests. (American Educational Research Association [AERA], APA, & National Council on Measurement in Education [NCME], 1999, p. 83)*

**Great Expectations! Choices, From SEA to CEE**

In 1988, the Government of Trinidad and Tobago appointed a task force (committee) to consider the removal of the CEE. The committee included several current and past educators along with a variety of prominent citizens representing various stakeholder groups, and was chaired by a former Minister of Education. The committee's terms of reference centred on the preparation of a plan for removing the CEE as the basis for placement of students in the secondary school (Trinidad and Tobago [T&T], 1988). Although this committee had the unique opportunity to

remove selective testing at 11+, it chose not to, arguing instead that there were secondary schools of varying quality that necessitated an adequate mechanism "to place the most academically capable students into the schools best equipped to maximise their potential" (p. 52). This vision of equity perhaps reflected loyalty to existing societal "folk norms," which maintain and legitimize elements of sponsorship in the selection process (London, 1989, p. 283). Surprisingly, as well, the committee frowned upon alternatives such as using scores from the continuous assessment and the zoning of students, both of which might have facilitated the removal or reduced the impact of a high-stakes one-shot examination.

In the end, the committee envisaged a selection procedure that, they believed, would (1) lessen the anxiety and stress associated with the CEE, (2) provide diagnostic and formative information on student performance and ability, and (3) ensure greater meritocracy. To accomplish these goals, the committee sought to redesign the selection instrument. The changes proposed included: (1) removing science and social studies, which the committee believed were not satisfactorily tested in the CEE; (2) increasing the mark for creative writing; and (3) limiting the instrument to achievement in English, mathematics, and creative writing, with an emphasis on reasoning and verbal ability.

The rationale for these decisions was founded upon a number of beliefs. For example, three reasons were given for restricting the choice of constructs. Firstly, it was argued that, in the past, components such as science, social studies, and creative writing exercised an inordinate influence upon the composite total score of the CEE. Secondly, it was believed that science and social studies were badly taught and inappropriately tested in the CEE. Thirdly, it was suggested that literacy and numeracy were the main goals of the primary school system and, therefore, low CEE scores in these areas were indicative of a system-wide problem. Concerning choice of format, it was argued that the MC format measured facts rather than critical thinking and facilitated guessing. At the same time, it was believed that CR tests were more authentic and better assessed higher-order thinking. A major expectation was for these changes to influence the pedagogical approach of teachers, encouraging a greater emphasis on teaching for critical thinking (Cheng, Watanabe, & Curtis, 2004).

Wisely, the committee also extended its analysis to the placement system, noting the lack of fairness. It was argued that the system was possibly biased against females, because it sought to artificially balance the number of boys and girls placed despite the disparity in the number of places and the superior performance of females. In terms of crafting a solution to this disparity, however, the committee believed that:

> It would be a simple matter to remedy this situation but one has to consider the possible consequences of placing students purely on the basis of merit without any consideration of balancing the selection on the terms of gender. (p.43)

Nevertheless, the committee mistakenly believed that placement issues were no longer relevant with the implementation of universal secondary education, and so did not consider the need for further reform in this area.

**Judging the choices: Evaluating the SEA test design**

From an assessment perspective, the faith of the committee in the ability of a one-shot examination to fairly select and allocate students to different life opportunities was surprising. Perhaps this credulity was reinforced by the perceived transparency of the selection process and a lack of trust in teachers' judgements (London, 1989). However, from a psychometric standpoint, it is obvious that multiple assessments would better capture both ability and achievement at 11+ (Henderson-Montero, Julian, & Yen, 2003). Indeed, the 1999 version of *Standards for Educational and Psychological Testing* emphasized the fallibility and limitations of single measures of achievement, especially in the context of high stakes testing. Notably, Standard 13.7 stated:

> In educational settings, a decision or characterization that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the validity of the decision. (AERA et al., 1999, p. 146)

This standard alludes to the superiority of multiple measures, whether in terms of multiple samples, different formats, or different measures across time (Henderson-Montero et al., 2003).

To a large extent, the committee's major focus was on the purpose or use of the selection instrument, with practicality and fairness considered to a much lesser extent. However, the full range of fairness issues was not explored and little consideration was given to the unintended consequences that might result from choice of format and construct. More than that, the committee failed to balance the expectation of positive impact with the likelihood of adverse consequences resulting from the choices made.

Admittedly, there was some support in the assessment literature for many of the arguments put forward by the committee, especially in its decision to make greater use of CR items. For example, Snow (1993) listed eight plausible rival hypotheses concerning the use of CR and MC items, and argued that CR tests were a better measure of ability and higher-order thinking and most likely promoted understanding and critical thinking. Likewise, Martinez (1999) summarized seven similar propositions about the use of CR and MC item formats and argued for an increased use of CR items, stressing that the range of cognitions elicited by extended CR items are usually broader than that assessed by MC item formats. He agreed that CR items might promote higher-quality learning through the washback effect.

However, the literature is not unified in its support for the exclusive use of CR item formats in a high-stakes test used for selection. Likewise, not everyone is euphoric about the value of CR items as measures of higher-order skills or as a mechanism to stimulate washback. Indeed, Mehrens (1998) has pointed out that there is little evidence to support the claim that the use of extended CR items in high-stakes testing impacts positively upon teaching and learning. Similarly, empirical studies on the psychometric difference between MC and CR items have not been conclusive. For example, Bridgeman (1992) found little difference between open-ended and multiple-choice formats in the Graduate Record Examinations (GRE). Similarly, Martinez (1991) compared stem equivalent figural MC and CR items and found only slight differences in statistical performance. In a later study of licensing tests in the field of architecture, even though figural CR items were more difficult, MC items proved more discriminating (Martinez & Katz, 1996).

A recent meta-analysis of past studies by Rodriguez (2003) confirmed that stem equivalent MC and CR items measure the same construct equally well. Moreover, it has been found that the use of CR items does not necessarily prevent test-takers from employing deleterious strategies such as working backwards from an answer (Katz, Bennett, & Berger, 2000). Hancock (1994) has explicitly tested the hypothesis that MC items cannot measure higher-order skills when compared with CR items. He found little support for differences in the two formats, but suggested the need for greater skill in the writing of MC items to test higher-order skills. Bearing in mind the weight of the evidence, some have even argued that in terms of testing time and cost-effectiveness, it might be better to replace a limited number of CR questions with a large number of MC items (Kennedy & Walstad, 1997).

**Administration, aftermath, and myth**

With hindsight, perhaps, there was no concerted effort to implement all the committee's recommendations in the test development stage. For example, it is unclear whether the committee's stated focus on higher-order thinking, development of diagnostic capacity, and tasks related to ability were actually translated in the development of the 2001 SEA. Nevertheless, prior to the administration of the examination, there proved to be tremendous support for the proposed design. This came from both prominent educators and writers in the popular press (Findlay, 2001; McDowall, 2000; Ragbir, 2000). Reasons given for the strong support centred upon the perceived focus on critical thinking, reduction in guessing and anxiety, and expectation of significant washback. Indeed some were suggesting that washback had already occurred (Allen-Agostini, 2001). On publication of the results, however, the possibility of adverse consequences soon turned into reality, with the then Minister of Education observing that "looking at the statistics yesterday of the SEA results and remembering the Common Entrance results, by far the majority of students who perform at low levels are in fact boys" (Pickford-Gordon, 2001, p. 5).

Consequently, the great majority of students assigned to "one-special" classes and schools were males, with 1,812 males to 704 females in 2001 and 2,240 males to 1,010 females in 2002. Such an adverse impact is significant, whether or not the reasons for it are construct or format related.

## The Design of the Study

### Research questions and hypotheses

This study was designed to determine the size and direction of gender differentials in performance and placement in the 2001 SEA. The study also investigated the possibility that gender differentials vary across schools, school types, and ability grouping. A variety of indices were used to analyze the data, including Cohen's d-values, SDRs, and F/M ratios.

The key research questions were:

1. What are the size and direction of gender differentials in overall scores and on each component of the SEA?

2. To what extent are there differential outcomes for males and females in the placement process?

3. To what extent do the size and direction of the differentials vary by school and school type?

4. To what extent do gender differentials and placement outcomes vary across ability groups as measured by total score SEA?

The literature provides support for the hypothesis that gender differentials will be small or negligible for mathematics, larger for language arts, and largest for creative writing (Garner & Englehard, 1999; Pomplun & Sundbye, 1999). Because the design of the SEA battery included two language components, the differential for the composite total score should strongly favour females (Willingham & Cole, 1997). It is possible that patterns of gender differentials in both performance and placement would either be uniform or vary across ability groups. Based on the findings in Jules (1994), however, the latter option is more likely. One possibility is for negligible or small differences among males and females in the high-ability groups, and larger differentials favouring females in the lower-ability groups. On the other hand, it might well be that high-achieving females will also have an advantage over high-achieving males on a test composed solely of CR items (DeMars, 1998).

In terms of gendered placement patterns, there are three possible options. One possibility is for equal placement opportunities for males and females. This is likely if males and females are allocated separately and/or the numbers of available places are equal. However, if males and females are competing for a specific number of places, then placement opportunities should favour females, with females more likely to receive their choice of school. A third possibility is for placement opportunities to favour males. This will occur if there are more "first and second choice" school places for males. In terms of the equity, one might argue that if parental choice is the sine qua non of the placement process, then both males and females should have comparable placement opportunities based on their stated choice of secondary school. On the other hand, it can be argued that, even with a flawed test design, placement opportunities should be solely dependent upon test scores regardless of gender.

**Sample and methodology**

The original data set consisted of 2,258 students in 44 schools in the St. George East Educational District (Smith, 2002). This district is reported as having one of the highest mean scores in the CEE (Jules, 1994). The 44 schools included 5 private and 7 single-sex primary schools. For this sample, the 7 single-sex primary schools, including two private schools, were excluded. The total number of schools in the sample was therefore 37, inclusive of 4 private schools. The total number of students was 1,896. In terms of the target population, the St. George East Educational District includes 61 public schools, 19 of which are government-run and 42 government-assisted (T&T, 2001). However, one of the new public schools included in the sample was not listed. Overall, then, the sample represented more than 53% of the schools in the district.

The profile of the schools in the sample is provided in Table 1. As shown, the sample included mainly government, Hindu, and Roman Catholic (RC) schools. The majority of schools were situated in the semi-urban areas along the East-West Corridor. However, the sample also included two rural RC schools. The Presbyterian, government, and Anglican schools were usually larger and the RC and Hindu schools smaller. The largest numbers of candidates in the sample were from the government and Presbyterian schools. Mean total SEA scores were highest for the

private and Presbyterian schools and lowest for the government and RC schools.

The data were analyzed using SPSS for Windows 9. For each school, means and standard deviations for the overall SEA score and each of the components were calculated. An EXCEL spreadsheet was used to calculate Cohen's d and the SDR. Schools in the sample were then categorized based on the size of Cohen's d. The seven categories and associated d-values were: 1) females performed much better (>0.8); 2) females performed better (0.5 to 7.99); 3) females performed slightly better (0.2 to 4.99); 4) males and females performed similarly (-0.199 to 0.199); 5) males performed slightly better (-0.2 to -4.99); 6) males performed better (-0.5 to -7.99); and 7) males performed much better (<-0.8).

**Table 1. Profile of Schools in the Sample (St. George East Educational District)**

| Denomination | | Location | | | Mean Size | | No. of Candidates | | Mean |
|---|---|---|---|---|---|---|---|---|---|
| Type | No. of Schools | Ur. | Semi Ur. | Rur. | Cap. | Pop. | Total M | F | SEA Score |
| Anglican | 2 | - | 2 | - | 672 | 660 | 107 | 80 | 618.17 |
| Gov't | 12 | 2 | 10 | - | 727 | 507 | 444 | 336 | 557.54 |
| Hindu | 6 | - | 6 | - | 388 | 273 | 105 | 104 | 589.47 |
| Muslim | 2 | - | 2 | - | 485 | 500 | 58 | 64 | 604.54 |
| Presb. | 3 | - | 3 | | 569 | 724 | 154 | 158 | 644.03 |
| Private | 4 | - | 4 | - | - | - | 36 | 38 | 646.53 |
| R.C. | 6 | - | 5 | 2 | 363 | 292 | 103 | 104 | 545.50 |
| SDA | 2 | 1 | - | 1 | 298 | 290 | 29 | 44 | 602.16 |

To obtain the percentage number of students placed according to choice for each gender, the six parental choices and the actual placing of the student were recorded. The percentage of students receiving each placement choice was then calculated. To obtain the female to male choice ratio for students at different ability groupings within the sample, percentiles were calculated using the students' composite total score. The percentile data was then used to create five ability groupings. The groupings were: 1) below the 30$^{th}$ percentile, 2) below the 50$^{th}$ percentile, 3) above the 50$^{th}$ percentile, 4) above the 75$^{th}$ percentile, and 5) above the 90$^{th}$ percentile. For each category, the weighted proportion of males and

females receiving choices 1 to 6 along with the Cohen's d and SDR was calculated.

**Results**

**What are the size and direction of gender differentials in performance and placement?**

Table 2 provides the means, F-ratios, p-values, SDRs, and Cohen's d-values for 37 schools in the sample categorized by denominational type. As shown by the means, p-values, and Cohen's d-values, females performed better on all three SEA components and on the composite total score. However, the differences were negligible for mathematics (0.193), relatively small for language arts (0.405), and medium-sized for creative writing (0.511). Overall, the standard mean difference for the composite total score (0.409) was relatively small. However, bearing in mind that composite total scores are used to determine placement, the size of this differential may be considered educationally significant, since it often results in different outcomes for males and females. When compared with the sample of CEE scores in a 1999 study analyzing students in the lower percentile who repeat the CEE, this Cohen's d-value was higher[1]. The SDR indicated that male scores were more widely distributed than that of females. This meant that at lower percentiles males were likely to be over-represented.

Table 3 includes the percentage of males and females placed according to the listed parental choice—1 to 6. As shown, males were more likely to receive placement choices 1 to 4. However, females were favoured on the lower placement choices 5 and 6, or were assigned to a school by the Ministry of Education. The chi-square value suggests that the overall distribution of the placement choices in this district was significantly different from what one would expect by chance. This implied that there were differences in placement opportunities for males and females. It is possible, however, that these particular placement patterns were district specific, dependent upon gender-based performance differentials, choice patterns, and the availability and distribution of schools.

**Table 2. Means, F-ratios, p-values, Cohen's d-values, and SDRs for the Three Major Components of the 2001 SEA (St. George East Educational District)**

| | Male | | Female | | | | |
|---|---|---|---|---|---|---|---|
| **SEA Components** | **Mean** | **SD** | **Mean** | **SD** | **p** | **Cohen's d** | **SDR** |
| Mathematics | 198.92 | 31.50 | 207.59 | 27.68 | 0.000 | 0.193 | 0.961 |
| Language | 195.73 | 32.01 | 208.47 | 27.06 | 0.000 | 0.405 | 0.956 |
| Essay | 194.45 | 31.92 | 209.76 | 27.65 | 0.000 | 0.511 | 0.957 |
| Overall | 589.09 | 89.95 | 625.81 | 76.09 | 0.000 | 0.409 | 0.961 |

**Table 3. Numbers and Percentages by Gender of Students Placed According to the Six Parental Choices (total data set)**

| **Choices** | **Students Placed** | | | | | | **F/M Placement Ratio** |
|---|---|---|---|---|---|---|---|
| | **Males** | | **Females** | | **Total** | | |
| | **No.** | **%** | **No.** | **%** | **No.** | **%** | |
| 1 | 210 | 17.7 | 157 | 14.6 | 367 | 16.3 | 0.825 |
| 2 | 160 | 13.5 | 110 | 10.2 | 270 | 12.0 | 0.755 |
| 3 | 178 | 15.0 | 125 | 11.6 | 303 | 13.4 | 0.733 |
| 4 | 203 | 17.1 | 130 | 12.1 | 333 | 14.7 | 0.707 |
| 5 | 194 | 16.4 | 184 | 17.1 | 378 | 16.7 | 1.042 |
| 6 | 124 | 10.5 | 181 | 16.9 | 305 | 13.5 | 1.609 |
| Total | 1,069 | 90.3 | 887 | 82.6 | 1,956 | 86.6 | 0.837 |
| Assigned | 115 | 9.7 | 187 | 17.4 | 302 | 13.4 | 1.794 |

N=2,258, $\chi$=65.065, df=6, p=.000

**Do gender differentials vary by school type & institution?**

Figure 1 shows the distribution of schools in the four categories constructed using Cohen's d. There were very few schools with medium or large gender differentials in favour of males. Indeed, only two schools fell in these two categories. However, in mathematics, there were 7 schools with small differentials favouring males, 8 schools with negligible differences, and 13 schools with small differentials in favour of females. Eight schools also reported negligible differentials for language arts and seven schools for creative writing. On the other hand, although only 3 schools reported medium-sized differentials favouring females for mathematics, as many as 11 schools reported such differences for language arts and eight for creative writing. Significantly, 8 schools reported large gender differentials favouring females in creative writing. This suggests that medium to large size differentials favouring females were more likely in language arts and creative writing than in mathematics. In terms of the composite total score, it may be significant that as many as 9 schools reported negligible gender differentials. Twelve schools, however, reported medium-sized differentials favouring females and 5 reported large differentials. This suggests that schools in this sample were more likely to report a female advantage in the SEA composite total score.

Schools grouped within the various gender differential categories had no unique identifying characteristics. However, schools reporting large gender differentials favouring males often sent up fewer candidates. Nonetheless, there were also schools with large numbers of candidates in which gender differentials for all three SEA components were either negligible or consistently small in favour of males. These schools were likely to report the highest SEA total scores. It follows that schools reporting medium-sized to large differentials in all three components tended to be either low achieving urban and rural schools. This suggested that a positive learning environment contributed to a reduction in the size of differentials, and that large differentials favouring females might reflect both overall underachievement and a poor learning climate.

Nevertheless, the expected pattern of negligible differences in mathematics, small differences in language arts, and medium-sized

differences for creative writing was not found in all schools. A few schools reported large differentials in specific SEA components while minimizing differences in others. In some cases, the component in which differences were minimized was unexpected. For example, in one semi-urban small Hindu school there were large differences favouring females in mathematics but medium-sized differentials for language arts and creative writing. Similarly, one semi-urban government school in the San Juan/El Socorro area reported medium-sized differences for language arts but small differentials for mathematics and creative writing. This suggests that differentials may have been created, in part, by institution-specific factors such as administrative support, the organization of instruction, the management of resources, learning climate, and teaching-learning focus.

It was unclear whether denominational grouping was an important variable influencing the size and direction of gender differentials. For example, in Hindu schools, two of six schools reported differentials in favour of males for all three components, whereas the other four schools reported large differences in favour of females. Individual Presbyterian, RC, and private schools were likewise found in the extreme categories. On the other hand, government schools were often found in the middle categories, reporting either low gender differentials in favour of females or negligible differences in all three components. The two Anglican schools in the sample were large in size and also reported large differentials in favour of males.

**A case study: Institutional factors vs. school type**

The three Presbyterian schools in the sample highlighted the weakness of using denominational type as a variable in the analysis of gender differentials. All three schools were situated in suburban areas and sent up large numbers of candidates. The schools were all high-achieving, with mean composite total scores above 620. However, differentials for mathematics, language arts, creative writing, and the composite total scores varied in both size and direction. For example, the highest-achieving school reported small mean differences in favour of males in mathematics (-0.210) and negligible differences for language (-0.032). The Cohen's d for creative writing (0.250) was small and in favour of females.

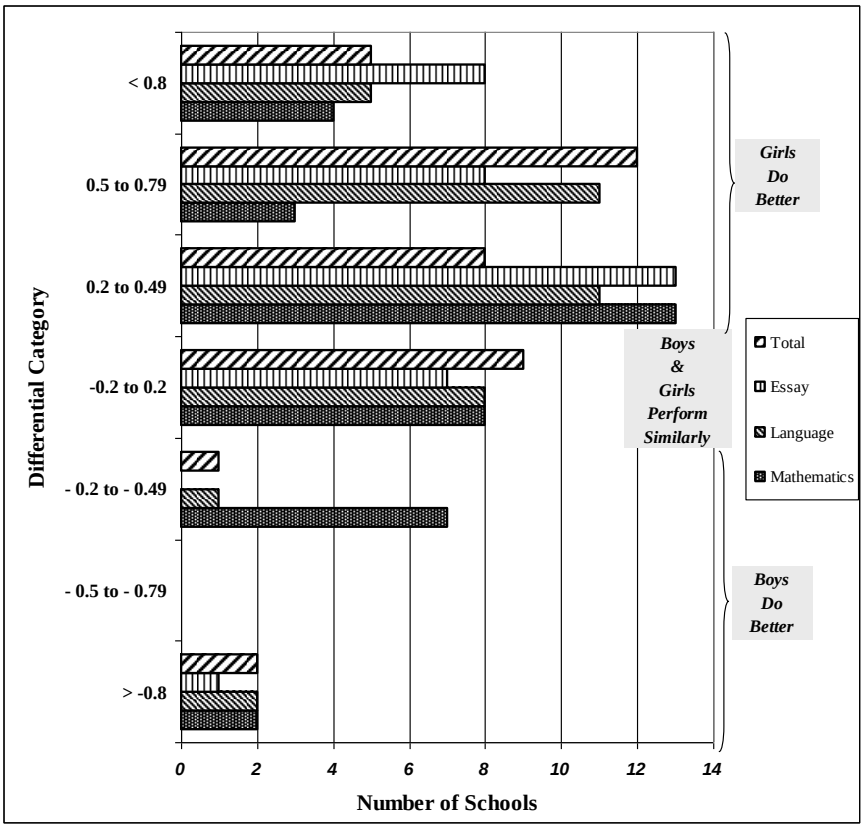


Figure 1. Number of schools in sample categorized by size and direction of differentials for scores on total SEA and components.

Significantly, in this school, the distribution of mathematics scores was more variable for females than for males (SDR=1.158). In the second school, all differentials were small and in favour of females, ranging from 0.245 for mathematics to 0.482 for writing. By contrast, in the third school, the differentials, all in favour of females, were more variable. Significantly, the differentials were small for mathematics (0.442), medium-sized for language arts (0.734) and large for creative writing (0.970). The SDR in all cases indicated that the distribution of scores for males was more variable than that for females. These findings confirm that institution-specific variables were more important in creating or reducing gender differences in achievement compared with overall school type or physical characteristics, such as size.

**Do gender differentials in placement and performance vary across ability groups?**

Table 4 provides the placement choice ratios, SDRs, and Cohen's d-values across different ability groupings. As shown, at all ability levels males were more likely to receive their first choice compared with females. The reverse was true, however, for students receiving the lowest choices (5 and 6) and those assigned to a school by the ministry. Below the 50th percentile, males were more likely to receive a favourable placing for choices 1 to 4. This was also true for students above the 50th percentile. However, this pattern of gender disparity was reversed for students above the 75th percentile, with females favoured on placement choices 2 to 6. Above the 90th percentile, males were favoured on their first choice. Notably, despite their high score, 36 females received their second choice and 13 received their third choice. This suggests that very high- and low-ability females were more likely to receive unfavourable outcomes in the placement system.

The Cohen's d for the composite total score was negligible for students in the top half of the class but approached a medium-sized effect in the lower half. This suggests that the greatest disparity was between low-achieving males and females, with differentials reduced in the higher-ability groupings. This pattern held for language arts, with medium-sized differentials favouring females for the lower half of the class reversed for students above the 75th percentile. In mathematics, for students above the 50th percentile, the magnitude of the differential was negligible and in the direction of males. Notably, as well, above the 75th percentile, differentials were also small and in favour of males. On the other hand, for creative writing, differentials were consistently in favour of females. For the lower half of the class, the Cohen's d was medium-sized whereas for students above the 50th percentile the differentials were small. Based on the SDR, the distribution of scores for males was restricted for language arts, creative writing, and the composite total score. Interestingly, above the 50th percentile in mathematics, female scores were more variable.

**Table 4. Female to Male Weighted Placement Choice Ratios, Cohen's d-values, and SDRs at Different Percentiles (for total score)**

| Choices | F/M ratio @ percentile | | | | |
|---|---|---|---|---|---|
| | ≤30th | ≤50th | ≥50th | ≥75th | ≥90th |
| 1 | 0.94 | 0.73 | 0.66 | 0.70 | 0.63 |
| 2 | 0.47 | 0.34 | 0.75 | 1.30 | *NM (36 F) |
| 3 | 0.42 | 0.50 | 0.85 | 1.16 | *NM (13 F) |
| 4 | 0.74 | 0.65 | 0.84 | 1.63 | *NM (1 F) |
| 5 | 0.93 | 1.03 | 1.48 | 1.34 | 0 |
| 6 | 1.33 | 1.46 | 3.35 | 3.28 | 0 |
| Total | 0.85 | 0.86 | 0.91 | 1.00 | 1.49 |
| Assigned | 1.70 | 1.73 | 11.83 | 0.73 | 0 |
| **Cohen's d** | | | | | |
| Total Score | 0.444 | 0.477 | 0.152 | 0.056 | 0.085 |
| Mathematics | 0.155 | 0.219 | -0.140 | -0.253 | -0.156 |
| Language Arts | 0.482 | 0.503 | 0.071 | -0.054 | -0.060 |
| Essay | 0.436 | 0.507 | 0.326 | 0.251 | 0.240 |
| **SDR** | | | | | |
| Total Score | 0.854 | 0.827 | 0.977 | 0.970 | 0.927 |
| Mathematics | 0.855 | 0.821 | 0.997 | 1.038 | 0.977 |
| Language Arts | 0.913 | 0.891 | 1.096 | 1.186 | 1.036 |
| Essay | 1.003 | 0.897 | 0.972 | 10.51 | 1.183 |

***NM (_F) = No Males (Number of Females Assigned)**

## Discussion

Overall, the results show that the pattern of gender differentials in placement and performance was complex and sometimes even contradictory, varying across school and ability level. The paradox is that while overall SEA composite scores favoured females, placement patterns appeared to benefit males more. These two contradictory and discriminatory patterns existed together, and therefore it might be premature to conclude that there is an advantage (or disadvantage) to any one group. The finding of different and opposing patterns in performance and placement supports the idea that the entire selection process should be studied when examining consequences (Messick,

1989). It appeared that the SEA test design, with its choice of construct and format, may have contributed to a measure of gender inequity in the system. However, the sizes of the differentials for the composite total score were relatively small, below the benchmark of 0.5 for a medium-sized effect. Nevertheless, it is possible that an alternative test design might reduce the size of this differential and improve the validity of inferences.

It was notable that the predicted pattern of gender differentials was not found in all schools, even within the same denominational type. Therefore, generalizations about the impact of school type on gender differences were not very useful. These findings confirm the minimal role of school type in the creation of gender differentials (Harker, 2000; Yang & Woodhouse, 2001). At the same time, there was evidence that institution-specific variables influenced the creation or reduction of gender differences. One such factor may be the quality of the learning environment (Evans, 2001). This is an issue that also needs further study.

The pattern of gender differences in placement choice may be considered an issue of distributive justice. These inequitable patterns were possibly brought about by differences in the availability of school places for males and females within the education district. Even with the implementation of universal secondary education, this source of inequity remained and therefore should be considered in the redesign of the selective system. The inequity in placement opportunities also provides opportunity for further study. Key questions to consider are, "*How can differences in placement opportunities be reduced?"* and *"What factors motivate the choice of schools and in what ways are the patterns different for boys and girls?"* While the placement system may have been designed to ensure that the top 20% receive their choice of school, there is little evidence that this does take either place consistently or fairly.

Some have suggested that gender differentials may vary across ability groupings; however, this issue has not been studied in the English-speaking Caribbean (Figueroa, 2002; Parry, 2000). It is significant, then, that the data in this study support this hypothesis. It is clear that the question is not necessarily "*Are boys underachieving?*" but "*Which boys are underachieving?*" (Epstein, Elwood, Hey, & Maw, 1998). In light of the contradictory patterns found, however, an additional question may be, "*What structures disadvantage either males or females in situations where they*

*should receive favourable outcomes in the selection procedure?*" The analysis of ability groups suggests that small to medium-sized gender differentials favouring females were more likely to be found at lower percentiles, with differentials negligible in the top half or last quartile. Creative writing was an exception, with differentials either small or medium-sized and always in favour of females. It may also be significant that female distributions were more variable at higher percentiles. While the overall findings support the hypothesis that "*not all males underachieve*," it also emphasizes the poor performance of males in the lower-ability groups, who appear to be doing significantly less well than females in the same ability range (De Lisle & Pitt-Miller, 2002). Indeed, the concern over the academic and social classroom performance of low-achieving boys in the classroom has been noted elsewhere:

> Low attaining boys frequently received negative attention from classmates and teachers, this was also characteristic of girls (to a lesser extent). This attention was not focused on all low attainers, it tended to be focused on just a few students. Low attaining students displayed poor basic school skills (such as reading) and poor social skills. Particularly among boys, evidence has been provided showing poor reading (especially reading aloud skills). These boys also did not show care or concern for classmates. They were responsible for class punishments and often teased their classmates. (Kutnick et al., 1997, p. 21)

Such poor social and academic skills will likely hinder remediation efforts in the "one special" classroom or special school.

We believe that the current debate on gender policies in schooling must move beyond the rhetoric of pro-feminist pedagogies on one side, and male recuperative philosophies and backlash politics on the other, to focus on ways in which schools can become efficient learning organizations for all students, with an emphasis on structures and policies for inclusion (Mills, 2000, 2003). We agree that gender differences must be seen as one of the conceptual keys to unlocking the range of discourses about "effective learning" (Daniels, Cresse, Hey, Leonard, & Smith, 2001; Fielding, Daniels, Creese, Hey, & Leonard, 1999). The work of Evans (2001) on streaming in Jamaican schools is illustrative of the approach needed. Indeed, she provides evidence of a common

organizational arrangement that may lead to the kinds of inequity found in this study.

We argue that using the 30% cut score on the SEA to place students in the "one special" classroom or school may increase the potential for misclassifications because low-ability males are more likely to do poorly on language-based CR tests, independent of their "true ability." For example, it is likely that some students (possibly male) would not have been classified as "one special" if either a MC or multiple measure test format was used (Chester, 2003; Kennedy & Walstad, 1997). The numbers of misclassifications (false negatives) may even be higher because the SEA measures fewer achievement constructs with fewer items. Additionally, the CR format does not provide response options, so students with very low language skills are less likely to provide an answer, thereby reducing the amount of information on performance available. In effect, the SEA may discriminate against low-ability males, without providing the diagnostic information promised.

This study has important implications for the design of high-stakes tests. It may be that the committee did not fully weigh the benefits against the costs in deciding on the choice of construct and format in the new test design. Additionally, the committee may have failed to consider alternative test designs that are gender-balanced and more congruent with the purpose of the assessment (Chilisa, 2000). We argue that a multiple measure test design is more likely to achieve this goal (Chester, 2003).

Five alternative test designs for the SEA are presented in Table 5. As shown, the most straightforward alternative is to include a MC-section in mathematics, measuring knowledge and problem solving through context-dependent items. This may provide a balance for the two language components and would better reflect an appropriate emphasis on numeracy, a construct equally important for success in the secondary school. A second option may be to add a language arts MC-component. This might work to reduce the differentials in both the component and composite total score. A third alternative would be to retain the social studies and science section, where gender differentials are likely to be negligible or small, combining both components into a test of general knowledge. Scores on this component would likely balance the higher

differentials in the language arts and creative writing, which now favour females.

**Table 5. Five Alternative Test Designs for the SEA**

| **Change in Test Design** | **Impact/Rationale for Change** |
|---|---|
| 1. Add MC component in mathematics | Improve validity of inferences for success on quantitative tasks in secondary school |
| | Reduce gender differential in composite score |
| 2. Add MC component in language arts | Reduce gender differential in composite score |
| 3. Add general paper MC component combining science and social studies | Improve validity of inferences for measure of achievement in primary school |
| | Reduce gender differentials in composite score |
| 4. Add CA component in science and social studies | Improve authenticity and provide a better measure of the construct |
| | Reduce gender differentials in composite score |
| 5. Add CA component in writing | Improve authenticity |
| | Reduce gender differentials in composite score |

The first three options are traditional approaches and do not make use of multiple measures across time, thereby limiting the validity of inferences. The third and fourth options, however, assess students continuously or at different times during Standards 4 and 5. In terms of writing, a more authentic design might be to allow students to engage in

the real process, from drafting and editing to the creation of an authentic final piece. While this approach will not reduce the size of gender differentials, it may be argued that the gender-differentiated skills assessed are both construct- and task-relevant. In terms of authenticity, we agree with the committee that the science and social studies constructs are better assessed holistically and in context. One possibility might be to use standard performance assessment tasks administered at a particular time in the lower forms. The tasks may be scored by a team of teachers across schools using well-developed rubrics, with scores moderated by a team of principals and supervisors in the area.

A decision to redesign a high-stakes test is an important one, critical not only to the life chances of examinees but also to the resolution of inequity and the future economic development of Trinidad and Tobago. Such decisions should never be made lightly and must be guided by the weight of empirical evidence. We believe that such decisions should minimize reliance on implicit beliefs and folk norms. Moreover, for increased transparency and assessment literacy, issues related to fairness and equity in the testing must be fully articulated in public. Indeed, it may be that the apparent meritocracy of the current selective system is more apparent than real and, in fact, only a few students gain an advantage, since most students are placed in the new-sector schools (either by choice or by the Ministry of Education). The situation may be as Chilisa (2000) reminds us:

> Assessment, especially when it takes the form of a national examination, is the most powerful tool that those who control the schools use to assert their power . . . . Whenever gender inequalities in academic achievement are observed, we ought to ask the following questions: How is achievement defined? Who defines achievement and in whose interest? What is the purpose of achievement? Who grades? Who defines the criteria for grading? What messages do the language, the content and materials used in the assessment tasks convey?" (p. 61)

## Note

1. The Cohen's d in this study, which was presented at the 1999 Biennial Cross-Campus Conference on Education held at the School of Education, UWI, St. Augustine, was 3.63 for the first score and 3.52 for the second score (N=616).

**Disclaimer**

This study was conducted with data used in the second author's M.Ed. Thesis. The data were reworked using the conceptual framework and analytical methods described. It was completed when the author was a student and is not in any way connected to his current employment.

**Acknowledgments**

The authors wish to thank Yvonne Lewis, Director of the Division of Educational Research and Evaluation (DERE) for providing the data; Rita Bhagwandeen, Cheryl Gomez, Nyla Abdool and the rest of the library staff at the School of Education, UWI, St. Augustine for help in locating the newspaper articles; Nicole Henry for her help with editing; and Stacy Quavar of the Centre for Medical Sciences Education, UWI, St. Augustine for reworking the data.

## References

Allen-Agostini, L. (2001, March 28). Minister says: SEA to reward merit – 22,000 take new test tomorrow. *Trinidad Guardian*, p. 23.

Alloway, N., & Gilbert, P. (1997). Boys and literacy: Lessons from Australia. *Gender and Education, 9*(1), 49-58.

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

American Psychological Association. (2001). *Publication manual of the American Psychological Association (5th ed.)*. Washington, DC: Author.

Bailey, B. (2000). School failure and success: A gender analysis of the 1997 General Proficiency Caribbean Examinations Council (CXC) examinations for Jamaica. *Journal of Education and Development in the Caribbean, 4*(1), 1-18.

Behrman, J. R. (1996). *Human resources in Latin America and the Caribbean*. Washington, DC: Inter-American Development Bank.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement history examination. *Journal of Educational Measurement, 31*(4), 275-293.

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement*, 29(3), 253-271.

Bridgeman, B., & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31*(1), 37-50.

Broadfoot, P. (2002). Beware the consequences of assessment! *Assessment in Education: Principles, Policy and Practice, 9*(3), 285-288.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds.). (2004). *Washback in language testing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Chester, M. D. (2003). Multiple measures and high-stakes decisions: A framework for combining measures. *Educational Measurement: Issues and Practice, 22*(2), 32-41.

Chilisa, B. (2000). Towards equity in assessment: Crafting gender-fair assessment. *Assessment in Education: Principles, Policy and Practice, 7*(1), 61-81.

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

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-220). Washington, DC: American Council on Education.

Daniel, L. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the Schools, 5*(2), 23-32.

Daniels H., Hey, V., Leonard, D., Fielding, S., & Smith, M. (1999). *Learning and gender: A study of underachievement in junior schools.* (ESRC Report R000237346). Swindon, UK: Economic and Social Research Council.

Daniels, H., Cresse, A., Hey, V., Leonard, D., & Smith, M. (2001). Gender and learning: equity, equality and pedagogy. *Support for Learning, 16*(3), 112-116.

De Lisle, J., & Pitt-Miller, P. (2002). Not all males underachieve! Evaluating gender-based differentials in academic achievement at a medical school. *Journal of Education and Development in the Caribbean, 6*, 87-110.

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*(3), 279-299.

DeMars, C. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55-77.

Elwood, J. (1999a). Equity issues in performance assessment: The contribution of teacher-assessed coursework to gender-related differences in examination performance. *Educational Research and Evaluation, 5*(4), 321-344.

Elwood, J. (1999b). Gender, achievement, and the 'gold standard': Differential performance in the GCE A level examination. *Curriculum Journal, 10(2)*, 189-208.

Epstein, D., Elwood, J., Hey, V., & Maw, J. (1998). Schoolboy frictions: Feminism and 'failing' boys. In D., J. Elwood, V. Hey, & J. Maw (Eds.), *Failing boys? Issues in gender and achievement* (pp. 3-18). Buckingham, UK: Open University Press.

Evans, H. (2001). *Inside Jamaican schools*. Mona, Jamaica: UWI Press.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research*, *94*(5), 275-283.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62(1)*, 61-84.

Fielding, S., Daniels, H., Creese, A., Hey, V., & Leonard, D. (1999). The (mis)use of SATs to examine gender and achievement at Key Stage 2. *Curriculum Journal, 10*(2), 169-187.

Figueroa, M. (2002). Making sense of the male experience: The case of academic underachievement in the English-speaking Caribbean. *YouWe Quality Assurance Forum, 8: 11-16*

Findlay, M. (2001 March 18). SEA of worries: Last-minute concerns over new exam. *Sunday Express*, p. 11.

Fuller, B., Hua, H., & Snyder, C. W. (1994). When girls learn more than boys: the influence of time in school and pedagogy in Botswana. *Comparative Education Review, 38*(3), 347-376.

Gallagher, A. M. (1997). *Educational achievement and gender: A review of research evidence on the apparent underachievement of boys.* Bangor, NI: Department of Education, Northern Ireland. (DENI Research Report; No. 6)

Garner, M., & Engelhard, G., Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items. *Applied Measurement in Education, 12*(1), 29-51.

Gipps, C., & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Buckingham UK: Open University Press.

Goldberg, N., & Bruno, R. (1999). Male *underachievement in Dominica: Extent, causes and solutions*. Roseau: Ministry of Education, Sports and Youth Affairs.

Gorard, S., Rees, G., & Salisbury, J. (1999). Reappraising the apparent underachievement of boys at school. *Gender and Education, 11*(4), 441- 454.

Gray, D., & Sharp, B. (2001). Mode of assessment and its effect on children's performance in science. *Evaluation and Research in Education, 15*(2), 55-68.

Harker, R. (2000). Achievement, gender and the single-sex/coed debate. *British Journal of Sociology of Education, 21*(2), 203-218.

Herbert, S., & George, J. (1996). A tracer study of student performance on CXC Chemistry and "A" level chemistry in Trinidad and Tobago. *Caribbean Curriculum, 6*(1), 1-21.

Henderson-Montero, D., Julian, M. W., & Yen, W. M. (2003). Multiple measures: Examination of alternative design and analysis models. *Educational Measurement: Issues and Practice*, *22*(2), 7-12.

Jules (1994). *A study of the secondary school population in Trinidad and Tobago: Placement patterns and practices*. St. Augustine: Centre for Ethnic Studies, UWI.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-mathematics items: It's not the strategy. *Journal of Educational Measurement*, *37*(1), 39-57.

Kennedy, P., & Walstad, W. B. (1997). Combining multiple-choice and constructed-response test scores: An economist's view. *Applied Measurement in Education*, *10*(4), 359-375.

Kutnick, P. (1999). Quantitative and case-based insights into issues of gender and school-based achievement: Beyond simplistic explanations. *Curriculum Journal, 10*(2), 253-282.

Kutnick, P., Jules, V., & Layne, A. (1997). *Gender and school achievement in the Caribbean*. London: Department for International Development.

Layne, A., & Kutnick, P. (2001). Secondary school stratification, gender, and other determinants of academic achievement in Barbados. *Journal of Education and Development in the Caribbean*, *5*(2), 81-101.

Leo-Rhynie, E. (1999). Gender analysis in educational policy and practice. *Journal of Education and Development in the Caribbean, 3*(1), 1-17.

London, N. A. (1989). Selecting students for secondary education in a developing society: The case of Trinidad and Tobago. *McGill Journal of Education, 24*(3), 281-291.

McDowall, H. (2000, July 8). In defence of the SEA. *Trinidad Guardian*, p. 15.

Martinez, M. E. (1991). A comparison of multiple-choice and constructed response figural response items. *Journal of Educational Measurement*, *28*(2), 131-145.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*(4), 207-218.

Martinez, M. E., & Katz, I. R. (1996). Cognitive processing requirements of constructed figural response and multiple-choice items in architecture assessment. *Educational Assessment, 3*(1), 83-98.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence*? Education Policy Analysis Archives,* 6 (13). Retrieved July 20, 2003, from http://epaa.asu.edu/epaa/v6n13.html

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Miller, E. (1991). *Men at risk*. Kingston, Jamaica: Jamaica Publishing House.

Mills, M. (2000). Issues in implementing boys' programmes in schools: Male teachers and empowerment. *Gender and Education, 12*(2), 221-238.

Mills, M. (2003). Shaping the boys' agenda: the backlash blockbusters. *International Journal of Inclusive Education, 7*(1), 57-73.

Moss, P. (1997). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6-12.

Mullins, M., & Green, D. R. (1994). Gender bias in licensure testing: Is it a problem? *Clear Exam Review, 5*(2), 19-21.

Murphy, R. J. L. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology, 52*(2), 213-219.

Parry, O. (2000). *Male underachievement in high school education in Jamaica, Barbados, and St Vincent and the Grenadines*. Mona, Jamaica: Canoe Press.

Payne, M. A., & Barker, D. (1986). Still preparing children for the 11+: Perceptions of parental behaviour in the West Indies, *Educational Studies, 12*(3), 313-325.

Pickford-Gordon, L. (2001, July 6). Girls beat boys again. *Newsday*, p. 5.

Pomplun, M., & Sundbye, N. (1999). Gender differences in constructed response reading items. *Applied Measurement in Education, 12*(1), 95-109.

Ragbir, L. (2000, July 22). CE vs SEA: The difference. *Trinidad Guardian*, p. 15.

Rampersad, J. (1999). Patterns of achievement by gender in school science. *Caribbean Curriculum, 7*(1), 37-50.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Smith, P. (2002). *The 2001 Secondary Entrance Assessment: An analysis of student performance in an educational district of Trinidad and Tobago*. Unpublished master's thesis. The University of the West Indies, St. Augustine.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes? *Educational Researcher, 31*(3), 25-32.

Trinidad and Tobago. Ministry of Education. (2001*). Summary of public primary schools in Trinidad and Tobago according to planning boundaries*. Port of Spain, Trinidad: Author.

Trinidad and Tobago. Task Force for the Removal of the Common Entrance Examination. (1988*). Report.* Port of Spain, Trinidad: Ministry of Education.

Warrington, M., & Younger, M. (2000). The other side of the gender gap. *Gender and Education, 12*(4), 493-508.

Willingham, W. W. (1999). A systemic view of test fairness. In S. J. Messick (Ed.), *Assessment in higher education: Issues of access, quality, student development, and public policy* (pp. 213-242). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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. 56-126). Mahwah, NJ: Lawrence Erlbaum Associates.

Witt, E. A., Dunbar, S. B., & Hoover, H. D. (1994). A multivariate perspective on sex differences in achievement and later performance among adolescents. *Applied Measurement in Education*, *7(3)*, 241-254.

Yang, M., & Woodhouse, G. (2001). Progress from GCSE to A and AS level: Institutional and gender differences and trends over time. *British Educational Research Journal, 27*(3), 245-267.

Younger, M., Warrington, M., & Williams, J. (1999). The gender gap and classroom interactions: Reality and rhetoric? *British Journal of Sociology of Education*, *20*(3), 325-341.