Caribbean Curriculum
Vol. 13, 2006, 91–129

DRAGGING ELEVEN-PLUS MEASUREMENT PRACTICE INTO THE FOURTH QUADRANT
The Trinidad and Tobago SEA as a Gendered Sieve[1]

Jerome De Lisle

> *"Group impact and educational impact are the quintessence of social consequence. They drag measurement practice into that fourth quadrant".* (Willingham, 2002, p.196)

This paper expands upon concerns expressed earlier in De Lisle & Smith (2004) about the relationship between Eleven-Plus test design and patterns of gendered achievement in Trinidad and Tobago. Following Willingham's (1999), evaluation protocol, it includes (1) a critical analysis of gender fairness issues, (2) an empirical evaluation of gendered impact, and (3) a consideration of proposals for resolving gender fairness issues. Datasets used in the analysis are from the 2001–2003 Secondary Entrance Assessment (SEA) and the 1998–2000 Common Entrance Examination (CEE). The results confirmed that the gender gap was significantly greater for the SEA compared with the CEE, and that both males and females were disadvantaged in different ways by the placement system. The results also confirmed the existence of medium-sized gender differentials across urban-rural educational districts, literacy constructs, and high-low ability groups. A recent proposal to change the way the composite score is calculated did little to reduce the overall female advantage. Moreover, misclassification rates for the current remediation cutscore set at 30% were relatively high. These fairness issues are not easily resolved, but suggest the need for evidence-based test designs, test validation studies, and a re-examination of the need for selection.

[1] For the period of this work, which extended over two years, I acknowledge the assistance of the Division of Educational Research and Evaluation. While I am fully responsible for all statements made, feedback from this department helped to improve the study.

Introduction

High-stakes certifying and selection examinations have always had an important role in the education-oriented societies of the English-speaking Caribbean (London, 1989; Payne & Barker, 1986). Taking pride of place is the Eleven-Plus examination, which is taken by every student at the end of the primary school cycle (Heyneman, 1987; Stanley-Marcano & Alexander, 1998). By governing access to secondary schools of different quality, performance in this examination becomes critical to later success at 16 and 18+ and, ultimately, to opportunities for upward social mobility (Jennings, 2001). The use of selection examinations as early as age 11 may be especially critical in situations where opportunities for educational advancement and social mobility are very limited (MacKenzie, 1989). On the one hand, it is possible that such early selection will enhance system efficiency by ensuring that only those students who are best capable can gain access to further opportunity (Durbrow, Schaefer, & Jimerson, 2002; Heyneman, 1987). On the other hand, sorting and weeding students using high-stakes tests in an elitist school system can achieve at best only pseudo-meritocracy and may limit opportunities for widening access (Gould, 1996; Hickling-Hudson, 2002; Sacks, 2001a; Siegel, 2004).

Examinations used for selection purposes shape education systems and influence institutional structures and outcomes (Drenth, van de Flier, & Omari, 1983; Greaney & Kellaghan, 1995; Kellaghan & Greaney, 1992). This is because the examination acts as a gatekeeper (or gateway), creating differential test-taker outcomes, and governing access to further educational opportunity (Sacks, 2001b; Zwick, 2002). While many school systems have primary gatekeeping mechanisms at 16 to 18+, in the Caribbean, the role of the Eleven-Plus examination as an early gatekeeper has been inherited from the colonial era (Jules, 1994; Payne & Barker, 1986). Nevertheless, its continued retention is only possible in a society completely at ease with concentrating "the most disadvantaged children in the least popular [and worst resourced] schools" (Edwards & Tomlinson, 2002, p. 30). The Common Entrance Examination (CEE) at the end of the primary school cycle acted as gatekeeper by either barring or allowing entry to secondary school. Changing the purpose and design of the examination would not substantially alter this purpose; however, with the implementation of universal secondary education, the emphasis might

become sorting students towards different school types, as Hickling-Hudson (2002, p. 572) has noted in the case of Jamaica.

I would argue that a more useful analogy for a selection examination in this role is a "sieve," which emphasizes the concept of separation rather than debarment (Handy, 1989). The purpose of a simple sieve is to separate constituents; a concept expressed in popular terms, such as "*separating the wheat from the chaff,*" which implies that the constituents have dissimilar value. Using this alternative analogy, initially, the CEE would have acted as both gatekeeper and sieve by (1) barring entry into secondary school for some students, and (2) separating or allocating the remaining students to different school types. On the institution of universal secondary education in 2001, however, only the sieve function was retained in the newly designed Eleven-Plus examination—the Secondary Entrance Assessment (SEA) (Trinidad and Tobago. Task Force for the Removal of the Common Entrance Examination [T&T. Task Force], 1998). Although all students could now access secondary schooling, Eleven-Plus test scores were still used to decide which students received particular opportunities. Arguably, a new layer of separation was added, with students below a cutscore of 30% automatically allocated to remedial classes, thereby further ensuring efficient "separation of wheat from the chaff."

Whether or not opportunities have changed is an issue for education sociologists. Test evaluators are instead concerned with both adequacy of the sieve function and consequences of test use (Kane, 2001). Social consequences are also important to the test evaluator because the technical aspect of testing can never be separated from the social and political functions (Messick, 1988). In this regard, Willingham (1999) proposed a useful three-part protocol for evaluating test fairness, which involves (1) identifying fairness issues, (2) assessing impact, and (3) resolving fairness status. The first stage requires an analysis of the full range of issues associated with differential performance. In this paper, both the sieve analogy and an argument-based approach to test validation are used to identify and analyze critical gender fairness concerns. In the second part of the paper, empirical evidence for gendered impact is gathered from Eleven-Plus data; and in the final section, an attempt is made to resolve the fairness issues.

Of course, dual gatekeeping and sieve roles operate for selection examinations elsewhere. For example, in the US, SAT scores are critical determinants of access to institutions of higher learning (Sacks, 2001b). Even though high school grade point averages may be better predictors of undergraduate performance, a low SAT score will either bar a student or result in placement within a less prestigious institution (Camara & Kimmel, 2005; Zwick, 2002). Still, we might expect the fallibility of an Eleven-Plus gatekeeping system to be significantly higher than that at 18+. Indeed, most gatekeeping systems are inherently fallible because inferences and decisions are usually based upon a single composite score. Limitations can also arise if there is no theoretical rationale for the construct that the test score is supposed to measure or insufficient evidence to support inferences that the test measures the construct well (Gardner & Cowan, 2005; Sacks, 2001b).

Fairness and Precision as Examination Myth

There are important hidden benefits to using an examination for making high-stakes decisions on educational opportunity within small states. For example, only an examination will ensure the successful creation of the myth that selection is fair and precise, even when this is conceptually or technically impossible (Gardner & Cowan, 2005). Indeed, locally, the myths of precision and fairness have so successfully been promulgated that Eleven-Plus scores are often considered sacrosanct. Moreover, such myths can be augmented and sustained indefinitely through a lack of transparency. Consequently, although selection examinations have retained such a critical role within Caribbean education systems, there has been little consideration of issues related to test fairness. Indeed, internationally, there are still only a handful of studies of the Eleven-Plus in the Caribbean region (Durbrow et al., 2002). In the absence of these studies, those considered as illuminators (journalists, assessment evaluators, and other academics who bring light to the issue of test quality and fairness) are well advised not to express strong faith in the validity of the selection process.

Thus the real value of the precision and fairness myths may lie in the negation or suppression of dissent or challenge by stakeholders, including illuminators (Kane, 2002; Linde, 2003). Perhaps, then, it is this astounding absence of illumination coupled with the lack of transparency

that is central to the continued legitimacy of the Eleven-Plus selection system (Gardner & Cowan, 2005; Jules, 1994). London (1989), for example, has argued that considerable sponsorship exists in the Eleven-Plus selection process of Trinidad and Tobago, implying that the system is not as meritocratic as many believe. However, while London focused primarily upon the influence of paid lessons, there are also differences in the quality and extent of test preparation and variations in opportunities to learn across educational districts and schools (Haladyna & Downing, 2004). Indeed, local schools vary greatly in quality, especially across the urban-rural divide (De Lisle, Smith, & Jules, 2005). Thankfully, however, pervasive and institutionalized cheating is not yet a feature of Eleven-Plus selection in Trinidad and Tobago, although it might well become so if high-stakes school-based assessments were to be implemented (Overdorf, 2004). So, while there are still some who might seek to sneak over the wall with the help of the Concordat[1], this paper is concerned only with the gender fairness of the actual Eleven-Plus examination. Put another way, the question is: *To what extent is the eleven-plus gate equally open to boys and girls?*

Designing (and Redesigning) the SEA

It would be a grave error to assume that any test is automatically gender fair and, further, that gender fairness is independent of test design. The belief that "test design does not matter but scores from the test do" remains a significant problem in current analyses of gendered achievement in the Caribbean. Many empirical studies of gendered achievement assume that assessment is a completely neutral procedure and pay little attention to the changing nature of assessments (Gipps & Murphy, 1994; Kutnick, 2000; Layne & Kutnick, 2001). However, changes in assessment design must be considered when evaluating trends across subjects or time (Bailey, 2000; Rampersad, 1999). For example, analyses of 'O' Level results might consider changes in the weighting and nature of school-based assessments because these changes will alter the size of the gender gap (Elwood, 1999, 2005).

The assessment cycle includes the processes of test design, development, scoring, and administration, with the first two steps concerned with the knowledge and skills measured by the examination and the instruments used for measurement. It is at these early stages that critical choices must

be made about the purpose of the test, what constructs are required to effect those purposes, choice of assessment format, particular item types used, scoring methods, and statistical characteristics of items (Willingham & Cole, 1997). It is also significant that these test design and development decisions are interwoven into subsequent stages, thereby enhancing the impact of these choices. There is overwhelming evidence that a change in construct, format, or test blueprint will significantly alter the validity of a test for different subgroups, especially gender (Gipps & Murphy, 1994; Ryan & DeMark, 2002; Willingham & Cole, 1997). Since different choices will result in vastly different outcomes for males and females, test design or development committees must reflect on decisions and make public the predicted consequences of their choices, well before implementing any proposed changes (Willingham & Cole).

Acknowledging that a relationship exists between test design and gendered achievement is not the same as arguing that the increasing female advantage is the result of assessment schemes becoming "feminized" (Elwood, 2005). This argument implies some unholy conspiracy among test developers. Rather, the focus is on the need to develop explicit rules for a fair test. For example, neither construct nor authenticity can be the sole concerns in test design. Indeed, as Gipps and Murphy (1994) admitted, even test designers' conception of authenticity in a subject area is open to challenge. This has certainly proved true in the case of physics, where in the past masculine definitions of acceptable content have led to a relative disadvantage for females (Hildebrand, 1996). Indeed, Willingham (2002) has argued that when alternative test designs are available and one group of examinees is likely to be severely disadvantaged on a specific design, the most valid test is the fairest.

Considering the gender fairness of a chosen test design is the essence of analysing the Eleven-Plus as a gendered sieve. Such a focus is important because the format and structure of this examination has been significantly altered, and further changes have been proposed in an attempt to redress gender inequalities in performance (De Lisle & Smith, 2004; T&T. Division of Educational Research and Evaluation [DERE], 2004; T&T. Task Force, 1998). In 1998, the committee charged with the task of removing the CEE explored the issue of the gender fairness in selection, and concluded that there was grave injustice in any system that attempted to place an equal number of boys and girls. They noted that

"the present situation, if one judges on the basis of performance, is one in which girls who deserve places on the basis of merit are deprived selection in favour of boys who did not perform well" (T&T. Task Force, p. 41). Of course, this statement assumes that this or any other examination can objectively measure "merit" (whatever merit might be), and that scores assigned to the approximately 1,000 disadvantaged girls represent an accurate estimation of "capability to succeed in the secondary school," which seems to be the test construct in question.

The SEA is very different to the CEE in the fundamental design areas of construct, format, and weighting of the composite score (T&T. Task Force, 1998). Whereas the CEE tested knowledge of five subject areas—Mathematics, Language Arts, Creative Writing, Social Studies, and Science—the SEA omits the latter two subjects. The Task Force reasoned that although it should remain an achievement test, the instrument could be redesigned to better measure students' reasoning ability and verbal skills. Thus, compared with the CEE, the SEA composite score is heavily weighted towards linguistic-verbal competence, primarily ways of knowing language taught by the school (Myhill, 2005). From the literature, it is apparent that this choice might put some groups at a great advantage, most likely females and students of higher socio-economic status. Additionally, the test development decision to use constructed response items only would give these groups a further advantage. Thus, major choices in test design and development of the SEA are likely to result in females doing relatively better in the SEA compared with the CEE.

The adequacy of sampling may not have been an important consideration during test development, since the SEA had just over 100 items compared to 190 items in the CEE. The 2003–2004 test specifications for the SEA provide details of the sampling for that period (T&T. DERE, 2002). Of the 50 Mathematics items, 20 measure number, 18 measure measurement and money, 8 measure geometry, and 4 measure statistics. Thus, only two of three content strands are adequately sampled. Similarly, in Language Arts, 22 of the 50 items measure grammar, but only 8 measure vocabulary/spelling, 5 measure punctuation, 10 measure comprehension, and 5 measure graphic representation. Thus, only one Language Arts content strand is adequately sampled. While there is limited information on the scoring of questions, for the Creative Writing component, a holistic

rubric and two markers are used. Each rater gives a score between 0 and 6 using the rubric, which is then combined for a final score, with a maximum of 12.

Associated with the implementation of the 2005–2006 SEA was a redesign proposal intended to reduce the effect of the essay on the composite score (T&T. DERE, 2004). This proposal was based on the belief that the weighting of 1:1:1 for the SEA components gave too much weight to the essay. Mistakenly, then, this proposal viewed the essay as a single item, although in fact this is an extended performance task scored with a six-point rubric. Interestingly, the proposal failed to deal with another more significant scoring problem—low discrimination caused by a sketchy holistic rubric. The problem of inadequate discrimination (most students get between 6 and 12 and differences are small) was first considered by the Task Force, which believed that the best solution would be to increase the essay score to 48. In the 2005–2006 redesign proposals, the decision was made to use a 10-point rubric (maximum score of 20 by two raters). At the same time, the impact of the essay will be reduced by weighting it 2:3 against the Language Arts section. Thus, the overall weighting of the components for the composite score would become 5:3:2.

Although these are significant changes, they are not referenced to the critical issues of validity, reliability, or fairness (T&T. DERE, 2004; T&T. Task Force, 1998). More importantly, in the context of this paper, Willingham and Cole (1997) and Willingham (2002) emphasized that evidence for the impact of alternative test designs must be obtained and compared *before* implementation. Comparative studies using measures of impact such as effect sizes, standard deviation ratios, and female to male ratios on valued outcomes will allow us to make judgements on the fairness of alternative test designs. The gendered impact of the proposed redesign is a significant issue because changes in gender differentials across educational districts might also influence placement opportunities. Therefore, rather than reducing the advantage to a particular group, such changes, when implemented, might result in unpredictable and spotty patterns of winners and losers among both males and females.

Validity as the Context for Evaluating Test Fairness

From the standpoint of an evidence-based assessment design, Mislevy, Steinberg, and Almond (2003) reminded us that measures of student learning are essentially machines "for reasoning about what students know, can do, or have accomplished, based on a handful of things they say, do, or make in particular settings" (p. 4). More importantly, assessments are "embedded in a cultural setting and address social purposes both stated and implicit" and "communicate values, standards, and expectations" (p. 4). This understanding of assessment is critical to making appropriate choices when designing examinations to act as efficient sieves. Firstly, as a machine, an assessment can be put together in several alternative ways, with some designs more efficient for specific purposes. Secondly, all assessments are limited by the representativeness of the sample and, therefore, the adequacy of the sample of items or tasks is a major design issue. Thirdly, the values, standards, and expectations embodied within the social purposes of an assessment cannot be divorced from the technical aspects. Thus, educational and social purposes are not opposing facets of some choice wheel but instead must be balanced. More importantly, the values, standards, and expectations embodied in the assessment must be explicitly stated and considered both in planning and in evaluation.

To put fairness concerns into context, we must focus, then, on the current meaning of test validity, which is the most important criterion for evaluating a test (Gersten & Baker, 2002). The central focus of current conceptions of validity is on test score use and interpretation (Kane, 2001). In the modern era, Messick put forward two important concepts that have influenced the process of obtaining validity evidence. Firstly, he proposed a unitary concept that envisions construct validity as the whole. Secondly, he proposed a four-facet model of evidence, consequences, test interpretation, and test use, in which the fourth quadrant focuses upon the social consequences of test score use (Messick, 1988, 1989). It is this fourth quadrant that requires evaluators to address questions of "unintended side effects" and impact when evaluating tests. This is an important goal because test evaluators are really public scientists working for the public good (Ryan & DeMark, 2002). Thus, evaluators must avoid either a conformationist bias or a tendency to single-mindedly defend

current testing programmes (Haertel, 1999), and instead "speak to a diverse and potentially critical audience" (Cronbach, 1988, p. 4).

Validating test score use and interpretation is a major activity for any test agency involved in high-stakes testing. Messick's (1989) definition of validity is the foundation of the validation process, as defined in the 1999 *Standards on Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Validation, however, remains a nebulous process because new ideas about validity are not easily applied. Indeed, there are currently no guidelines for determining the relevance or types of validity evidence that should be collected (Kane, 2004a). Recently, in response, some authors have put forward an argument-based approach in which validation becomes the process of constructing and evaluating arguments for and against proposed test use and interpretation (Haertel, 1999). A validity argument will provide an overall evaluation of the plausibility of proposed interpretations and use of test scores. According to Kane (2002, 2004a), the first step is the construction of *an interpretive argument*, which are statements laying out the assumptions and inferences about the test score and its proposed uses. The act of constructing an interpretive argument is critical because test evaluators often set out to judge a test without establishing first what these proposed interpretations might be (Kane, 2004b). Using this approach, test validation involves "obtaining and weighing evidence to support or refute the interpretive argument" (Haertel & Herman, 2005, p. 2).

Since the interpretive argument provides an explicit statement of the proposed interpretations or test use, it demands further evaluation, generalization, and extrapolation. To aid the construction of an interpretive statement, Kane (2002) distinguished between *descriptive* and *decision-based* interpretations. The latter is important when judging test use and consequences and must be justified by gathering a variety of evidence from different sources (Ryan & DeMark, 2002). Kane also extended the framework to include *semantic inferences*, focused upon what the test score means, and *policy inferences*, which involve the adoption of decision rules. Using Kane's (2002) framework, four semantic (1-4) and two policy inferences (5-6) related to SEA as a gendered sieve are listed below:

1. Males and females perform similarly on the tasks outlined in the test blueprint.
2. Performance on the range of test objectives is similar for males and females.
3. Achievement on the SEA test objectives provides a comparable indication of underlying ability for males and females.
4. Achievement on the SEA test objectives provides comparable secondary school placement opportunities for males and females.
5. The implementation of the SEA will lead to increased gender equity.
6. The adjustment in weighting will lead to increased gender parity.

These inferences suggest that a fundamental assumption of the SEA is that males and females should perform similarly on the same set of knowledge and skills, and gender equity and parity remain important policy goals.

The SEA as a Gendered Sieve

Although the concepts of fairness and validity are analogous, fairness issues must be explicitly addressed when evaluating tests (Stobart, 2005; Willingham 2002). Test fairness may apply to either individuals or groups. For the individual, test fairness is directly equivalent to validity whereas group fairness is comparable validity for different sub-groups (Willingham, 1999). When considering the SEA as a *gendered sieve*, the focus is on group fairness, meaning that the SEA might have different predictive validities for males and females. As with the SAT, this implies that the SEA functions differently for males and females and is a better predictor of outcomes for one group (Zwick, 2002; Zwick & Schlemer, 2004). Technically, if construct under-representation and construct irrelevant variance were absent, groups might perform similarly. In practice, however, there are always differences between sub-groups and therefore policymakers must consider *how large these differences are* and *how existing educational policy is altered*, especially for high-stakes tests. When analysing group impact, gender, ethnicity, social class, and geographic location are common focal points because social equity and justice demands fair distribution of opportunities across these dimensions.

It is likely that the changes of construct and format from the CEE to the SEA altered the magnitude of the gender gap and created differential placement opportunities for males and females (Murphy, 2000). Any

judgement of fairness in this context must be tempered by the values and expectations implicitly or explicitly espoused about the test and its purpose, as contained in the semantic and policy inferences. If a society strives for gender equity and excellence, it is hypocritical to choose an Eleven-Plus test design that accentuates gender differences. Likewise, it would be foolhardy to propose or implement changes that might further magnify gendered impact. It is clear, then, that in the Caribbean setting, test designers must always consider gendered impact when making policy decisions (De Lisle & Smith, 2004).

Fairness and the Complexities of Selection at Eleven-Plus

While the test is central to the selection process, there are other important elements to consider when analysing fairness (Messick, 1989). For example, both achievement scores and placement opportunities should be evaluated. Interestingly, the Task Force (1998) believed that any disadvantage to females would be fully resolved on the institution of universal secondary education; however, it did not consider the vagaries of the placement system. It might be that the placement system remains as a unique and separate source of inequity (De Lisle & Smith, 2004; Jules, 1994). This inequity is difficult to evaluate using solely a quantitative approach because different individuals will place different values on similar outcomes. For example, one parent might be pleased with a child passing for a less prestigious school, while another parent, whose child had the same composite score, becomes distraught over the outcome because they expected placement in a more prestigious institution (Gersten & Baker, 2002). However, there are some issues that can be answered through quantitative data analysis, including: *To what extent did the change in Eleven-Plus test design alter placement opportunities for males and females?*

Placement opportunities are partly dependent upon the composite score, choices made by parents, and the availability of schools in different geographical regions (Jules, 1994; T&T. Task Force, 1998). For example, if a parent were to choose only prestigious schools, which are also chosen by parents of other higher-achieving students, then the likelihood of that student being placed is significantly reduced. In turn, parental choice is itself determined by a myriad of factors such as the school's perceived prestige, accessibility, and geographical location (Bagley, Woods, &

Glatter, 2001; Ball, Bowe, & Gewirtz, 1995; Parsons, Chalkley, & Jones, 2000). For example, the current distribution of schools will result in fewer realistic opportunities for most rural students, who are located far away from the urban centres, where most highly prestigious schools are located.

While the perceived prestige of a school is primarily based on academic performance, parents will consider other factors when choosing schools, such as whether the school is single-sex or coeducational, religious affiliation, and sixth form availability (Hospedales, 1982). The 2003 data show that parents selected very different schools for choices 1 to 3. In fact, there are 29 different schools in the 36 possible places, suggesting that different geographic communities consider different schools as viable first, second, or third choice opportunities. Indeed, only one school, Queen's Royal College, was chosen in two choice categories. Additionally, while there are similar numbers of single-sex boy and girl schools in choices 1 and 3, choice 3 is dominated by mixed schools. This pattern might influence placement opportunities outside the first two choices, as would a strategy of parents choosing less prestigious schools as their first choice.

The situation becomes even more complex when we consider that the more prestigious the institution, the greater the number of parents selecting it for first choice. This would significantly decrease opportunities for access by students with lower scores. Table 1 provides data on the students choosing the 12 most prestigious schools and those who actually obtain their preferred first choice placement. While the interquartile range suggests wide variation, the mean scores indicate that most parents choose accurately. Nevertheless, considering that over 700 students chose each of the first four ranked schools, with only about 100 students gaining access to each (including the 20%), the chances of obtaining a first choice placement to a prestigious institution seems rather difficult. The difficulty is illustrated in the case of the top-ranked school, St. Augustine Girls, where the median score of those choosing is 677, just below the minimum entry score of 682. This is likely the score of a student gaining access through the 20% Concordat provision. At the same time, the minimum score of 628 for Naparima, the second ranked school, suggests that the Concordat rule may be applied flexibly and differently across like institutions.

Table 1. ***On the Chances of Obtaining the Students' First Choice*:**
Characteristics of Students Choosing & Obtaining Placement in the Top 12 Ranked Schools, as Measured by Performance in the 2003 SEA

Rank	Name of School (Gender Composition)	Characteristics of Students											
		Choosing School as First Choice							Obtaining School of Choice				
		No.	Mean	*SD*	Median	IQ range	Max	Min	No.	Mean	*SD*	Max	Min
1 (077)	St. Augustine Girls (G)	869	663.22	56.90	677	73.00	735	426	108	723.11	11.246	735	682
2 (053)	Naparima Girls (G)	782	674.22	51.60	690	66.25	738	460	117	722.86	13.336	738	628
3 (069)	Queen's Royal College (B)	729	620.10	66.50	631	90.50	731	359	107	707.59	9.494	731	696
4 (109)	St. Stephen's College (B)	707	610.55	67.44	619	94.00	728	377	120	704.60	17.709	728	616
5 (061)	Presentation College, Chag. (B)	661	636.73	75.39	654	104.50	739	379	83	722.58	13.983	739	654
6 (029)	Hillview College (B)	641	644.98	60.66	657	82.50	734	343	109	714.06	18.634	734	621
7 (037)	Holy Faith Convent (G)	602	633.11	67.84	645	95.00	734	335	108	713.23	17.325	734	621
8 (113)	North Eastern College (M)	577	550.48	67.31	556	103.00	704	332	148	667.57	11.689	704	650
9 (009)	Bishop Anstey High (G)	561	633.45	62.38	646	79.50	735	365	106	706.83	16.149	735	646
10 (065)	Presentation College, San'do (B)	554	657.01	56.65	668	71.00	737	347	111	718.32	10.937	737	685
11 (049)	Naparima College (B)	527	661.14	55.99	672	68.00	736	369	103	715.29	22.185	736	606
12 (097)	St. Joseph Convent, St. J. (G)	522	656.22	51.84	667	69.00	733	279	114	712.86	17.495	733	629

KEY: B — Boys G — Girls M — Mixed

As expected, the mean scores of those choosing top-ranked schools are lower for males. Therefore, although females may score higher, access to the prestigious single-sex schools may be restricted. Reasonably, any change in the number of single-sex schools for males or females will also significantly alter placement opportunities. Therefore, the patterns reported by Jules (1994) may have changed significantly with the secondary school building programme initiated by the Secondary Education Modernization Programme (SEMP). These complexities will result in inequity, especially in instances where some parents do not have a full understanding of the placement process or make poor choices.

Lessons on Fairness from the Geography of Gendered Inequity

One of the complexities that requires special attention is the differences across geographic regions. In some of the work on gendered achievement, it is often assumed that gender differences are homogenous and differences in geographical location are not important. Even when there is an understanding that only some males are underachieving, little thought is given as to where these males might be located. One common view, for example, is that laddish behaviours and anti-school masculinities are more common in urban areas influencing gender differences in achievement (Archer & Yamashita, 2003; Jackson, 2003). It does seem appealing to argue, even without evidence, that the laddish behaviours of males in some urban areas are the prime determinant of male underachievement. However, while this argument is instinctively attractive, other possibilities exist. For example, it might be that some girls are also significantly underachieving in these urban centres (DeBlase, 2003). Another possibility is that some rural males and females are especially at risk because of poor quality teaching-learning, community expectations, and socialization practices (Chevannes, 2001; De Lisle, Smith, & Jules, 2005). Therefore, determining variations in gendered achievement patterns across geographical locations is important in seeking explanations for why some males and females underachieve.

That large differences exist across education districts is neither a new nor startling revelation. Most recently, data from the 2004 national tests results confirmed that the gender gap is indeed larger in the South-Eastern and North-Eastern educational districts and in Tobago (De Lisle, Smith, & Jules, 2005). Lowered performance in these regions is probably

a consequence of multiple factors, including the ineffectiveness of schools, lack of resources including trained teachers, and the quality of teaching-learning. The geography of gender inequity should considerably lessen our faith in the fairness of selection examinations, especially the Eleven-Plus yardstick. For example, if the gender gap varies so significantly across regions, just how meritocratic can the selection process be? Is it really achievement being measured or just opportunities to learn? In such a circumstance, how efficient is the sieve at separating students of different ability or gender for remediation?

Principles of Fairness

A biased sieve is unfair because it allocates "too many" of a favoured group to valued outcomes and creates "underachievement" in the other. With hindsight, then, it is a great paradox for any society to construct general principles of fairness without first re-evaluating its use of Eleven-Plus selection (Schrag, 2004). Again, implicit in such an approach is the belief that examinations are automatically a fair selection tool. However, while no examination is perfectly fair, all assessments can be made fairer (Stobart, 2005). Indeed, Gipps and Murphy (1994) urged test designers to work continuously towards the goal of fairer assessment by paying greater attention to purposes, nature, administration, and scoring. Moreover, Willingham and Cole (1997) provided clear prescriptions for choosing different test design options. More specifically, both Chilisia (2000) and Stobart considered the issue of fairness within multicultural societies. Stobart, for example, emphasized that fairness and equity are issues of judgement and not just a matter of equal numbers. Chilisia noted that in multicultural systems the content and format of an examination raises the issue of whose knowledge is tested, and whether that knowledge is really of such great value. Thus, implicit in this issue is whether test designers are privileging certain types of knowledge in order to maintain a dominant culture. It is clear that large-scale assessment systems in multicultural societies such as Trinidad and Tobago must work towards the fairest possible test design fairness for gender and other grouping factors.

It is interesting to note that principles of test fairness have been published, including the *Code for Fair Testing Practices in Education*, which complements similar guidelines in the fields of licensing and credentialing (Council on Licensure, Enforcement, and Regulation &

National Organization for Competency Assurance, 1993; Joint Committee on Testing Practices, 2004; Zieky, 2002). These codes emphasize the different elements that contribute to fairness at various stages of the assessment cycle. The question is: *How does the SEA stand up to the principles in these codes*? Analysing the SEA on the 22 criteria proposed by the Joint Committee on Test Practices, most of the current SEA weaknesses centre on the failure to provide evidence for the current interpretation of test scores. If an examination is so important, then surely there should be greater efforts at collecting validation evidence, including studies of impact and fairness (Benson, 1998).

In Search of Evidence

Thus, the key gender fairness issues in Eleven-Plus testing are test design, variations in gendered achievement patterns, magnitude of gender differences, and differences in placement opportunities. This study will primarily consider the gendered impact of different Eleven-Plus test designs, including proposed changes to weighting the SEA composite score and use of the 30% remediation cutscore. Variations in gendered achievement patterns and differences in placement opportunities across regions are also analysed. Since there is no standard for determining the magnitude or impact of a gender difference, one approach might be to compare gender differences across the different design options and regions. In quantifying these differences, an effect size measure is essential to judging practical significance (Cohen, 1988; Schagen & Elliot, 2004; Thompson, 2002; Willingham and Cole, 1997). The magnitude (effect size) of the gender gap would allow us to judge the impact of the SEA as a gendered sieve. Following Willingham and Cole and Willingham (1999), standard deviation ratios and male to female outcome ratios were also used to measure adverse impact.

The five research questions guiding the study are:

1. How do gender differentials compare across the 1998–2000 CEE and 2001–2004 SEA administrations?
2. What was the pattern of gendered achievement and placement opportunities in the 2003 SEA?
3. How do gendered achievement patterns vary across construct, ability group, and regions in the 2003 SEA?

4. What is the gendered impact of changing the rules for computing the composite total score, based on the 2003 SEA data?
5. What is the accuracy and gendered impact of using an arbitrary 30% cutscore to select students for remediation, based on the 2003 SEA data?

Methods and Procedures

The main database was the 2003 SEA results for all 20,671 students taking the examination in that year. This database included raw and standard scores for each component and the composite score. The database was coded only for district, religion, class group, school, gender, age, number of attempts, and six parental choices. The Division of Educational Research and Evaluation (DERE) also provided selected data from the 1998–2000 CEE and the 2001 and 2002 SEA. In Microsoft EXCEL, new composite scores and rankings were created using the proposals put forward for the weighting and scoring scheme in the examination (T&T. DERE, 2004). To obtain the new weighting scheme of 5:3:2 for the Mathematics, Language Arts, and Creative Writing components, weights were applied directly to the standard score. This new standard score was then prorated to ensure equivalency, assuming an original standard score of 600±30. Alternative methods of calculating the standard score were also attempted, including (1) applying the weights directly to the raw scores, and (2) combining the raw scores of English and Creative Writing in a ratio of 3:2 and then obtaining a combined standard score.

In this study, both *r* (measures of association) and *d* (standardized mean differences) families of effect sizes were used to measure gender impact (Kline, 2004; Walker, 2004; Willingham & Cole, 1997). The *d* group effect size was the Cohen's *d* and the *r*-group effect size was eta-squared. The formula used to calculate the Cohen's *d* was $d = M_1 - M_2 / \sigma_{pooled}$, where M_1 is the mean for males, M_2 the mean for females and σ_{pooled} the pooled standard deviation. Eta-squared is the proportion of the total variance attributed to an effect and is generated directly from the output of SPSS, version 12.0. It is formally defined as the ratio of the effect variance (SS_{effect}) to the total variance (SS_{total}). Two rubrics were used for

making a qualitative decision on the size of the difference. For Cohen's *d*, the benchmarks used for assessing the size of the impact are 0.2, 0.5, and 0.8 and for eta-squared 0.1, 0.3, and 0.5. Odds and risk ratios were also used to measure relative impact in group distribution. The odds ratio was defined as the ratio of odds for an event in one group divided by the odds in another group. The odds ratio may be viewed as an effect size measure for categorical variables (Fleiss, 1994; Kline 2004). While there were no available benchmarks for making a qualitative inference on magnitude, both the odds and risk ratios allow assessment of the relative proportion assigned to each group (Grissom & Kim, 2005).

Impact data were obtained from a specially constructed Microsoft EXCEL database, with formulae inputted for calculating the standardized mean difference, confidence intervals, and female to male and standard deviation ratios (Thompson, 2002). Differences in effect size were calculated across educational districts and five ability groups based on percentile data. These groupings were (1) below the 25^{th} percentile ($\leq 25^{th}$), (2) below the 50^{th} percentile ($\leq 50^{th}$), (3) above the 50^{th} percentile ($\geq 50^{th}$), (4) above the 75^{th} percentile ($\geq 75^{th}$), and (5) above the 90^{th} percentile ($\geq 90^{th}$). These percentile ranges corresponded to the lower and upper quartile, the lower and upper two quartiles, and the first decile. Based on the different raw scores, students were ranked in the EXCEL database using the rank function and the different rankings were compared. Changes of ranks across different percentiles were also calculated as a measure of impact. For the CEE and SEA data that covered the periods 1999–2002, mean scores and standard deviations provided by the Ministry of Education were plugged into the effect size spreadsheets.

In judging the utility and impact of the current 30% cutscore, classification accuracy was considered equivalent to consistency (Young & Yoon, 1998). Rates of misclassifications were therefore compared to (1) the criterion-referenced cutscore and (2) a hypothetical cutscore that assumes equal performance of males and females. To obtain the criterion-referenced cutscore, a standard setting panel of 12 teachers was convened (Hurtz & Hertz, 1999). From the two-week process, nine usable forms were obtained. The panellists were all first-year undergraduate students pursuing the Bachelor's in Education or Bachelor's degree in Mathematics and Education. All students were

scholarship recipients and had taught for 2 to 10 years prior to attendance at the training colleges. Five had substantial experience at the fifth standard level, four were male, two were from Tobago, and one successfully completed the Measurement and Evaluation elective at training college. The group underwent extensive training in standard setting and received materials related to the 2003 SEA, including question papers and table of specifications.

Panellists were provided with rating forms, guidelines, and original papers related to standard setting. The original design of the standard setting protocol had to be altered because item scoring schemes or rubrics were not available. Therefore, instead of assigning specific item weights, all questions were weighted equally, with task demands incorporated in the probability judgements. Both probability and extended Angoff[2] procedures were used, which required panellists to either estimate the probability of the minimally competent student getting each answer right or select the most likely score (Cizek, Bunch, & Koons, 2004; Hambleton & Plake, 1995). The extended Angoff procedure was used for setting standards on the essay paper. The procedure involved analysis and discussion of the test and test blueprint, development of specific descriptors for each performance level, individual rating, and discussion of group results (Brandon, 2004). The intense discussion during the process was necessary to ensure consensus and accuracy of judgement (Hurtz & Auerbach, 2003).

Four levels of performance were considered, which demanded three separate decisions for each subject area (Cizek, Bunch, & Koons, 2004). The performance levels corresponded to the definitions constructed for the 2005 Trinidad and Tobago national tests. The focus of this study was on the last two categories, at which the pass/fail boundary is assumed to exist. Level 2 was defined as *"marginal academic performance, with work approaching but not yet reaching satisfactory performance, indicating only a partial understanding and limited display of skills required."* This level is sometimes labelled as "Partially Proficient," "Borderline," or "Nearly Meets the Standard of Work Required at this Level." Level 1 was defined as "*inadequate academic performance indicating little understanding and minimal display of required skills, with a major need for additional instructional opportunities, remedial assistance, and/or increased student academic commitment to achieve at*

the Proficient Level." This level is sometimes labelled as "Below Basic," "Well Below Proficient," or "Well below the Standard of Work Required at this Level." Students at this performance level would normally be assessed for remediation.

Results

1. How do gender differentials compare across the 1998–2000 CEE and 2001–2004 SEA administrations?

Table 2 shows the changes in the standardized mean difference (Cohen's *d*) for each subject component across three CEE and four SEA administrations for the period 1998–2003.As shown, gender differences were negligible to small for Mathematics across all seven cohorts regardless of design (Range of Cohen's d = 0.169–0.274). The gender gap was also consistently negligible for Science (Range of Cohen's d = 0.127–0.176) and negligible to small for Social Studies (Range of Cohen's d = 0.171–0.233). In contrast, the gender gap for performance on the Creative Writing component was medium-sized for five of eight cohorts (Range of Cohen's d = 0.466–0.585). The greatest change in differentials across the CEE and SEA occurred in the Language Arts component where the standardized mean difference varied from 0.208 to 0.336 for the CEE, but was consistently over 0.4 for the SEA (Range of Cohen's d = 0.417–0.436). This difference was likely a consequence of changes in format and content across the two examinations, including the removal of the more "gender-neutral" components of Science and Social Studies (Cohen's d = 0.290 [1998] to 0.340 [2000]) compared with the CEE (Cohen's d = 0.391 [2004] to 0.421 [2001]). While this change is small, it may be notable because the composite score is used to determine placement (Olejnik & Algina, 2000).

Table 2. Changes in Component and Composite Scores for the CEE & SEA Eleven-Plus (1998—2004)

Year	Exam	Numbers		Cohen's *d*						*SDR*					
		Male	Female	Maths	Science	Lang Arts	Social Studies	Essay	Comp. Score	Maths	Science	Lang. Arts	Social Studies	Essay	Comp. Score
1998	**CEE**	13,961	14,665	0.17	0.13	0.31	0.21	0.51	0.29	0.94	0.94	0.90	0.91	0.88	0.90
1999	**CEE**	7,976	8,408	0.19	0.15	0.28	0.17	0.55	0.29	0.95	0.94	1.00	0.96	0.88	0.94
2000	**CEE**	7,914	8,470	0.22	0.18	0.34	0.23	0.59	0.34	0.97	0.94	0.95	0.91	0.89	0.92
2001	**SEA**	8,168	8,216	0.27		0.44		0.47	0.42	0.92		0.91		0.90	0.90
2002	**SEA**	8,095	8,289	0.24		0.43		0.51	0.42	0.93		0.88		0.88	0.89
2003	**SEA**	8,188	8,196	0.22		0.42		0.51	0.41	0.89		0.80		0.84	0.84
2004	**SEA**	10,338	10,476	0.20		0.43		0.48	0.39	0.91		0.86		0.86	0.88

Table 3. Patterns of Gendered Achievement on 2003 SEA Subject Components Across Five Ability Groups

SEA Construct	Total Population				*Indices of gendered achievement across different ability groups*				
	Male		Female		≤25th	≤50th	≥50th	≥75th	≥90th
	Mean	*SD*	Mean	*SD*	---------- Cohen's *d* ----------				
Mathematics (100)	56.62	26.61	62.31	23.78	0.15	0.19	-0.13	-0.20	-0.23
Language Arts (100)	59.02	26.11	68.83	20.85	0.56	0.49	0.20	0.19	0.11
Creative Writing (12)	7.17	2.85	8.50	2.42	0.58	0.57	0.34	0.33	0.20
Composite Total (212)	122.71	53.56	139.64	45.05	0.46	0.40	0.05	0.03	-0.04
	P	Cohen's *d*	*SDR*						
Mathematics	0.00	0.23	0.89		0.86	0.87	1.05	1.07	1.11
Language Arts	0.00	0.42	0.80		0.89	0.83	0.92	0.94	0.96
Creative Writing	0.00	0.50	0.85		0.88	0.86	0.92	0.87	0.87
Composite Total	0.00	0.34	0.84		0.85	0.83	1.00	1.00	0.99

Changes in the standard deviation ratios were small in both subject component and composite scores for the CEE and SEA. However, again the most notable differences across the seven administrations and two test designs were in Language Arts (Range of SDR_{CEE} = 0.904–0.948; Range of SDR_{SEA} = 0.855–0.911) and total composite score (Range of SDR_{CEE} = 0.902–0.924; Range of SDR_{SEA} = 0.836–0.904), with smaller values indicating that the distribution of males is increasingly more variable. This suggests that more males were likely to be found in the tails of the distribution for the SEA composite score. Again, while this is a relatively small change, the overall impact of these differences will be notable when the 30% cutscore is applied.

2. What was the pattern of gendered achievement and placement opportunities in the 2003 SEA?

Table 3 provides data on the pattern of gendered achievement in each component of the 2003 SEA. Apart from the mean raw scores for males and females, indices of statistical and practical significance are shown along with standard deviation ratios. The table shows that while all differences were statistically significant, only differences on Creative Writing were medium-sized or practically significant (Cohen's *d* [*Creative Writing*] = 0.503). This effect size indicates a notable difference in performance between males and females. However, the standard deviation ratio for Language Arts showed that a greater number of males were located in the tails of the distribution (SDR [*Creative Writing*] = 0.799).

Table 4 captures the gendered impact of placement by providing data on the mean, maximum, and minimum score of males and females obtaining their six choices. In this instance, eta-squared was the measure of practical significance used for continuous data and both risk and odds ratios for categorical outcomes (placement received).

Table 4. Gender Differences in Composite Total Score and Percentage of Males and Females Obtaining Choices 1 to 6

Parental Choices	Gender	Mean	*SD*	Max	Min	Median	Eta ^{2*}	%	Chi ^{2}	P-value	Risk Ratio	Odds	Odds Ratio
								% Male and Female Obtaining Choice					
First	F	673.75	70.92	738	367	703	0.03	14.9	4.96	0.03	0.93	0.18	0.92
	M	644.05	97.45	739	368	685		16.0				0.19	
Second	F	650.51	73.96	723	369	674	0.05	11.6	0.00	1.00	1.00	0.13	1.00
	M	611.24	93.58	722	367	636		11.6				0.13	
Third	F	629.08	73.10	720	375	653	0.03	12.0	0.17	0.68	0.99	0.14	0.98
	M	603.88	82.87	715	365	631		12.2				0.14	
Fourth	F	617.21	69.58	713	370	638	0.03	11.9	11.11	0.00	0.89	0.14	0.87
	M	589.57	77.91	715	370	612		13.4				0.16	
Fifth	F	606.40	64.31	710	381	615	0.06	13.1	0.70	0.40	1.03	0.15	1.04
	M	570.35	77.05	693	366	578		12.7				0.15	
Sixth	F	592.69	58.45	702	376	598	0.09	15.7	5.39	0.02	1.08	0.19	1.09
	M	553.29	71.00	689	367	563		14.5				0.17	
None	F	575.71	64.74	698	367	589	0.10	20.9	5.17	0.02	0.98	3.80	0.92
	M	530.87	70.75	695	365	534		19.6				4.11	

*****All p < .001

Table 5. Changes in Ranks Based on New Weighting of Composite Score as Applied to the 2003 SEA Data

% Change in Rank*	Magnitude of Change by Rank	No. of Students Changing Rank						No. of Students Changing Rank by Direction of Change				Overall Change in Rank		
		Top 100	Quartiles				Lower 30%	Positive		Negative		F	M	F/M ratio
			1st	2nd	3rd	4th		M	F	M	F			
< 1	**1-199**	100	1,774	896	1,062	2,787	3,051	1,786	1,779	1,667	1,282	3,453	3,061	1.13
1-4.9	**200-999**		2,808	2,952	3,350	2,271	3,010	2,289	3,582	3,219	2,295	5,508	5,877	0.94
5-9.9	**1,000-1,999**		556	1,132	784	110	195	412	924	863	384	1,275	1,308	0.97
10-19.99	**2,000-3,999**		24	121	40	2	7	40	99	37	11	77	110	0.70
Overall % Change		82	99.65	100.00	99.94	99.65	99.81							
F/M Ratio	**Old Score**	1.17	0.55	1.01	1.20	1.51	0.61							
	New Score	1.17	0.58	1.04	1.21	1.38	0.64							

*No recorded changes greater than 20%

As shown, differences between achievement scores for females and males were statistically significant for all six choices and for students assigned to a school by the Ministry of Education. However, as judged from the values of eta squared (Range of eta^2 = 0.025–0.099), these differences were all negligible in terms of practical significance. Likewise, the value of chi-square was statistically significant for the fourth and sixth choices, indicating that relatively more males received their fourth choice and relatively more females received their sixth choice. Nonetheless, the magnitude of this impact was relatively small as measured by the odds and risk ratios (Range of odds ratios = 0.870–1.094; Range of risk ratios = 0.855–1.080).

3. How do gendered achievement patterns vary across construct, ability group, and regions in the 2003 SEA?

Table 2 also provides data on the standardized mean difference and standard deviation ratios across five ability groups: (1) below the 25th percentile, (2) below the 50th percentile, (3) above the 50th percentile, (4) above the 75th percentile, and (5) above the 90th percentile. As shown, gender performances in the top quartile were different to that in the lower quartile. In the top quartile, the difference in Mathematics (Cohen's *d* = -0.199) and Language Arts (Cohen's *d* = 0.191) was small. The difference between males and female scores on Creative Writing was also small but relatively larger (Cohen's *d* = 0.325). The consequent difference in the composite score was also negligible (Cohen's *d* = -0.34). For the lower quartile, only the difference in Mathematics achievement scores was small (Cohen's *d* = 0.150); whereas gender differences on Language Arts (Cohen's *d* = 0.560), Creative Writing (Cohen's *d* = 0.569), and the composite score (0.457) were all close to or above the benchmark of a medium-sized effect.

Thus, males in the lower quartile were primarily the ones who were underachieving compared with females in the same ability grouping. Moreover, this underachievement was primarily on measures of literacy. Males in the upper quartile and top 10% clearly do equally well in the SEA, and in the case of the top 10%, the gender gap on Mathematics actually showed a small advantage to males (Cohen's *d* = -0.230). The standard deviation ratios for the top and lower quartiles also indicated significant differences, with male scores more variable on all three

components in the lower quartile, but in the upper quartile, female scores were more variable (SDR = 1.109). However, even in this ability group, male scores were still more variable in Creative Writing (SDR = 0.867).

Table 4 provided data on the pattern of gendered achievement across districts and changes in the female to male ratio across the six choices and assigned group. The data suggested that both the gender gap and placement opportunities varied substantially across districts. In Tobago, for example, females had a small- to medium-sized advantage on the three SEA components, whereas in Victoria the differences on the components were negligible to small. Thus, the gender gap was relatively larger in low-performing districts such as Tobago and in the rural South Eastern and North Eastern regions. Overall, males were more likely to receive their first and fourth choice; however, there were also significant differences across educational districts. In Victoria, for example, males were more likely to receive their first choice (F/M ratio = 0.892), whereas in Port of Spain and Environs the chances are about equal (F/M ratio = 1.045). Males were severely disadvantaged on choices 1 to 4 in the Caroni region (F/M range = 0.707 [3rd choice] to 0.922 [1st choice]). Females also appeared especially disadvantaged on choices 4 to 6 in Port of Spain and Environs and St. George East. In both Victoria and Tobago, females were close to 1.5 times more likely to receive their second choice.

4. What is the gendered impact of changing the rules for computing the composite total score, based on the 2003 SEA data?

Table 5 provides data on the numbers changing rank in different parts of the score distribution when the new weighting scheme is applied. The changes resulting from the new scoring regime were relatively small, with 121 students in the second quartile changing rank, amounting to a 10–20% change. As shown, in the top 100 students, most changes were relatively small with no overall change in the female to male ratios. However, in the lower 30%, relatively more females were found under the new scoring regime, although again the magnitude of the change is relatively small, of the order of 0–5%. The greatest change occurs in the fourth quartile where relatively more males were positioned. Table 6 also provides details of the changes in ranks by gender. Although changes in

the weighting regime were designed to reduce the gender differences, males were still more disadvantaged, with most of the larger negative changes (decrease in rank) occurring for this group. This is a notable effect considering that the numbers of male and female students with changes are overall quite similar. It is also evident, however, that the new scoring regime will improve the position of males in the first (F/M [old scoring] = 0.55 & F/M [new scoring] = 0.58) and fourth quartiles (F/M [old scoring] = 1.51 & F/M [new scoring] = 1.38). At the same time, more males will be placed in the remedial class assuming application of the 30% cutscore (F/M [old scoring] = 0.61 & F/M [new scoring] = 0.64).

Table 6. Misclassification Rates for the Current Norm-Referenced Remediation Cutscore

Cut-score	Numbers of Candidates			F/M Ratio	Misclassification Rates for Equal Percentages		Misclassification Rates for a CRT Cutscore of 40	
	Total	F	M		F	M	F	M
5	164	145	19	0.13	-38.41	+38.41	-23.37	+23.37
10	522	432	90	0.21	-32.76	+32.76	-17.71	+17.71
15	913	732	181	0.25	-30.18	+30.18	-15.13	+15.13
20	1,408	1,079	329	0.30	-26.63	+26.63	-11.59	+11.59
25	1,986	1,466	520	0.35	-23.82	+23.82	-8.77	+8.77
30*	2,600	1,839	761	0.41	-20.73	+20.73	-5.68	+5.68
35	3,418	2,311	1,107	0.48	-17.61	+17.61	-2.57	+2.57
40 º	4,220	2,745	1,475	0.54	-15.05	+15.05	0.00	+0.00
45	5,220	3,256	1,964	0.60	-12.38	+12.38	+2.67	-2.67
50	6,375	3,852	2,523	0.65	-10.42	+10.42	+4.62	-4.62
55	7,481	4,388	3,093	0.70	-8.66	+8.66	+6.39	-6.39
60	8,826	5,062	3,764	0.74	-7.35	+7.35	+7.69	-7.69

º Levels 1/2 cutscore of 85/212 (Combined raw score) equivalent to 40.09%

5. What is the accuracy and gendered impact of using an arbitrary 30% cutscore to select students for remediation, based on the 2003 SEA data?

Table 6 shows the misclassification rates for the 30% cutscore in two situations: (1) a hypothetical situation in which the numbers of males and females are always equal and (2) the obtained criterion-referenced cutscore of 40. As shown, assuming hypothetical equality of male and female performance at each score point, misclassification rates decrease significantly at higher percentiles. This pattern occurs because of

differences in both score distribution and variability for male and female sub-populations, with male scores having a lower mean and higher variability at lower percentiles. As a result, there were many more males at lower percentiles. The norm-referenced 30% cutscore chosen by the Ministry of Education was well below the standards-referenced cutscore of 40%, which resulted in as many as 1,620 students being misclassified or categorized as not requiring remediation. This number included 906 females and 714 males. Thus, females were at a greater relative disadvantage because of the inaccurate determination of the cutscore (Relative misclassification rate = -5.68 %).

Discussion

This study was designed to evaluate the gender fairness of the Eleven-Plus. Using an evaluation protocol developed by Willingham (1999), gender fairness issues were first identified and analysed using an argument-based approach to test validation. This process led to the development of semantic and policy inferences on gender fairness of the SEA. This first stage was followed by an evaluation of the relative impact of the different CEE and SEA designs and variations in gender achievement patterns across educational districts. The findings of this stage confirmed that there were notable variations in gender differences for the Eleven-Plus composite score across different test designs, possibly because of (1) changes in the item format and (2) exclusion of gender-neutral subject components. There were also notable variations in the gender gap across educational districts. Based on the evaluation protocol, it is now left to resolve some of the fairness issues identified and evaluated.

The evidence suggested that tinkering with the system to alter the weighting of the three SEA components from 1:1:1 to 5:3:2 will only slightly improve the ranking of some high- and low-scoring males, but will not reduce the female advantage for the great majority of students. In fact, females in the middle parts of the distribution will gain a further advantage. These unexpected results may be explained by substantial differences in the shape and spread of score distributions for males and females. The data showed that the distribution of male scores were more platykurtic (flatter at the top) but less negatively skewed (fewer higher scores) (Male Distribution: Mean = 583.3, *SD* = 89.86, Kurtosis = -0.659,

Skewness = -0.425; Female Distribution: Mean = 617.0, *SD* = 75.9, Kurtosis = -0.139; Skewness = -0.638). Thus, male and female distributions on the SEA composite score were very different in terms of both score location and variability, with more females located in the centre (the female distribution is taller at the centre), and increasingly more males found in the lower tail (Halpern, 2000). Thus, the larger number of females in the centre of the distribution reduced the utility of the new changes in component weighting.

The key questions in this resolution phase is whether the SEA should be modified or is currently acceptable for fair use and, if used, what constraints should be put in place. There is no single answer to these questions because the goals and purposes of a selection system cannot be resolved through using or modifying a single examination. Indeed, rewarding accomplishment, improving efficiency, and ensuring equal opportunity are competing objectives that will always result in contradictory choices. Three general approaches to resolving fairness issues are to: (1) remove existing barriers, (2) implement a compensatory approach to mitigate disadvantages, and (3) adopt a more democratic approach emphasizing formative testing and individual development. While the former two approaches have been tried in the past, the continued fallibility of Eleven-Plus testing suggests a need for considering the last approach.

Perhaps, the weakness of the SEA as a gendered sieve was most evident in the sharp variations in gendered achievement patterns and placement opportunities across the eight educational districts. If SEA scores were a fair and equivalent separator for males and females, then large differences in achievement scores and placement opportunities across regions should not be so readily apparent. However, this is not the case and there were very large differences in favour of females in the rural South Eastern and North Eastern regions, and on the island of Tobago. Of course, one might explain this finding by arguing that most males in this region were simply of lower ability. However, it is more likely that limitations in the instrument coupled with the variations in the quality of instruction and opportunities to learn were the main factors contributing to lowered performance among some males. It is also possible that differing definitions of male achievement and success are part of the complex of factors that work to create larger differentials in these rural communities.

These impacts may be more notable on some constructs, such as literacy (Chevannes, 2001; Connell, 1995, 2000; Rowan, Knobel, Bigum, & Lankshear, 2002).

As suggested earlier by Jules (1994), the placement system operated quite differently to the achievement component and was a separate source of inequity, especially for some regions and placement choices. Females certainly appeared slightly disadvantaged on the highly desired first choice placement, most likely because of the very high prestige (desirability) of one or two all-girl secondary schools and the performance of the females in the top ability group. Males were more likely to receive their lower choices in Port of Spain and Environs, probably because of the greater number of school places in this district. However, in Caroni educational district, males were severely disadvantaged on the first four choices, as there were fewer high-achieving, all-male schools in this region and spaces in the co-educational schools were also limited.

Resolving fairness issues also requires greater attention to the processes of test design and test evaluation. For test design, one solution might be to decide what type of gatekeeping or sieve role is required, what type of selection system will achieve the desired purpose, and then collect evidence that proves a particular design is comparatively fairer and more valid than alternatives. Indeed, with full universal secondary education implemented, it just seems easier to focus on improving the quality of secondary schools currently considered less prestigious. This will certainly reduce the need for a selection system to allocate students to different school types. Indeed, even if a fairer examination could be designed, just how useful would it be to track students towards differential opportunities at such an early age? Nevertheless, a fairer Eleven-Plus examination is possible. For example, a meaningful composite score can be achieved through multiple measures spaced over time, format, and construct (Goldberg & Roswell, 2001; Schafer, 2003). This is by no means a plug for the inclusion of high-stakes school-based assessment because this will raise the additional question of, *how exactly do we maintain fairness and integrity for teacher-administered school-based assessments when such high-stakes are involved?* (Murphy, 2000) The answer to this new question cannot lie in trusting teachers more, because the issue at hand is about conflict for highly valued, sparsely distributed resources in a society that has never come to grips with fairness and

meritocracy. Therefore, an externally administered and managed examination might be the only legitimate option at this time.

Assessment experts must pay greater attention to the principles of evidence-based assessment design, which is based on three premises:

1. Assessments should be built upon important knowledge, skills, and attitudes and focus on how these competencies are acquired and put to use.
2. The line of reasoning from what test-takers say and do, and inferences of what they know, can do, or should do next must be based on the principles of evidentiary reasoning.
3. The test purpose should be the driving force behind all design decisions, reflecting constraints, resources, and conditions of use.

According to Mislevly, Almond, & Lukas (2004), the three core design questions are:

1. What complex of knowledge, skills, or attitudes should be assessed?
2. What behaviours or performances reveal these constructs?
3. What tasks or situations might elicit these behaviours?

One benefit of such an evidenced-based approach is that the design process becomes systematic and tasks are matched to key outcomes. Additionally, test developers should be required to evaluate the impact of newly proposed designs prior to full implementation. Clearly, more evaluation information is required to support current test use and interpretation associated with the Eleven-Plus. Certainly, there is little evidence in this study to support the continued use of a 30% cutscore. Moreover, there is no theoretical rationale or educational criteria for applying such a cutscore, especially when the evidence indicates that SEA operates differently for lower-ability males and females. It is also questionable whether useful diagnostic information is available from the current SEA design (Green & Weir, 2004). As such, even with a criterion-referenced "remediation" cutscore, a properly designed and developed diagnostic test must be applied afterward. Indeed, the act of misclassifying potentially remedial students has grave implications for constructing a fairer selection system.

(Re)organizing to Find the Evidence

Relevant departments in the Ministry of Education must be retooled and better resourced to ensure that, as a public agency, evidence is collected for current interpretations and test score use. This would require a fuller range of assessment and psychometric services, including procedures and personnel to design, develop, and evaluate assessments. The evaluation role is critical, as there are currently insufficient validity studies of high-stakes, large-scale assessments used for placement or accountability purposes. The role of the higher education institutions is also important in supporting such an agency. This requires improvement and updating of programmes in the area of measurement theory. Cooperation between different institutions is essential if national goals are to be achieved and if evidence-based decisions are to be used to guide further development of the SEA.

Considering the current role of high-stakes examinations in the Caribbean, stakeholders must receive more information about important issues associated with the use of test scores. Blanket acceptance of test scores as prescriptive of a child's future is outdated. The role of illuminators, then, becomes central to good measurement practice and greater consideration of these issues is required. Academic illuminators must obtain the tools to study large-scale assessments and prescribe ways in which testing can become fairer. In hindsight, then, there seems little advantage in moving from the CEE to the SEA. There is little available evidence that the SEA has greater predictive validity or can adequately measure or promote critical thinking and problem solving (T&T. Task Force, 1998). Instead, the evidence from this study suggests that the SEA is considerably less gender fair than the CEE. It seems incredible, then, that such a fallible examination might continue to have such a critical role sieving males and females towards future opportunities at such an early age.

In Memory of

Omesh Matura, 2004 Winner of the President's Medal for Teachers and 2005/2006 Year 1 B.Ed UWI, St. Augustine student; a potentially great measurement student who was involved in this study's standard setting exercise.

Notes

[1] The Concordat is a formal agreement between Government and the Denominational School Boards. One element of this agreement is that 20% of the Eleven-Plus places assigned to a denominational school are reserved for nominees of the Board, subject to criteria laid out by the Ministry of Education.

[2] Standard setting is the process used to obtain a passing score for an examination. The Angoff procedure is one of the more widely used test-based methods.

References

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

Archer, L., & Yamashita, H. (2003). Theorising inner-city masculinities: 'Race', class, gender and education. *Gender and Education, 15*(2), 115–132.

Bagley, C. A., Woods, P. A., & Glatter, R. (2001). Rejecting schools: Towards a fuller understanding of the process of parental choice. *School Leadership and Management, 21*(3), 309–325.

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

Ball, S., Bowe, R., & Gewirtz, S. (1995). Circuits of schooling: A sociological exploration of parental choice in social class contexts. *Sociological Review, 43*(1), 52–78.

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practices, 17*(1), 10–17.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17,* 59–88.

Camara, W. J., & Kimmel, E. W. (Eds.). (2005). *Choosing students: Higher education admissions tools for the 21st century.* Mahwah, NJ: Lawrence Erlbaum.

Chevannes, B. (2001). *Learning to be a man: Culture, socialization, and gender identity in five Caribbean communities.* Mona, Jamaica: UWI Press.

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

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Connell, R. W. (1995). *Masculinities.* Sydney: Allen & Unwin.

Connell, R. W. (2000). *The men and the boys*. Berkeley, CA: University of California Press.

Council on Licensure, Enforcement, and Regulation, & National Organization for Competency Assurance. (1993). *Principles of fairness: An examining guide for certification boards*. Lexington, KY: Author.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 3–17). Hillsdale, NJ: Lawrence Erlbaum.

DeBlase, G. L. (2003). Missing stories, missing lives: Urban girls (re)constructing race and gender in the literacy classroom. *Urban Education, 38*(3), 279–329.

De Lisle, J., & Smith, P. (2004). Reconsidering the consequences: Gender differentials in performance and placement in the 2001 SEA. *Caribbean Curriculum, 11*, 23–55.

De Lisle, J., Smith, P., & Jules, V. (2005). Which males or females are most at risk and on what? An analysis of gender differentials within the primary school system of Trinidad and Tobago. *Educational Studies, 31*(3), 393–418.

Drenth, P. J. D., van de Flier, H., & Omari, I. M. (1983). Educational selection in Tanzania. *Evaluation in Education, 7*(2), 93–217.

Durbrow, E. H., Schaefer, B. S., & Jimerson, S. (2002). Diverging academic paths in rural Caribbean village children: Predicting secondary school entrance from the St. Vincent Child Study. *School Psychology International, 23*(2), 155–168.

Edwards, T., & Tomlinson, S. (2002). *Selection isn't working. Diversity, standards and inequality in secondary education* (Catalyst working paper). London: Catalyst Forum.

Elwood, J. (1999). 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. (2005). Gender and achievement: What have exams got to do with it? *Oxford Review of Education, 31*(3), 373–393.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

Gardner, J., & Cowan, P. (2005). The fallibility of high-stakes '11-plus' testing in Northern Ireland. *Assessment in Education 12*(2), 145–165.

Gersten, R., & Baker, S. (2002). The relevance of Messick's four faces for understanding the validity of high-stakes assessments. In G. A. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 49–66). Mahwah, NJ: Lawrence Erlbaum.

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

Goldberg, G. L., & Roswell, B. S. (2001). Are multiple measures meaningful?: Lessons from a statewide performance assessment. *Applied Measurement in Education, 14*(2), 125–150.

Gould, S. J. (1996). *The mismeasure of man* (revised & expanded). New York: Norton & Norton.

Greaney, V., & Kellaghan, T. (1995). *Equity issues in public examinations in developing countries* (World Bank Technical Paper No 27). Washington, DC: World Bank.

Green, A. B., & Weir, C. J. (2004). Can placement tests inform instructional decisions? *Language Testing, 21*(4), 467–494.

Grissom, R. J., & Kim, J. J. (2005*). Effect sizes for research: A broad practical approach*. Mahwah, NJ: Lawrence Erlbaum.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Haertel, E. H., & Herman, J. L. (2005). A historical perspective on validity arguments for accountability testing. *Yearbook of the National Society for the Study of Education, 104*(2), 1–34.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17–27.

Halpern, D. F. (2000). *Sex differences in cognitive abilities (3rd ed.).* Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41–55.

Handy, C. (1989). *The age of unreason.* Boston, MA: Harvard Business School Press.

Heyneman, S. P. (1987). Uses of examinations in developing countries: Selection, research, and education sector management. *International Journal of Educational Development, 7*(4), 251–263.

Hickling-Hudson, A. (2002). Re-visioning from the inside: Getting under the skin of the World Bank's Education Sector Strategy. *International Journal of Educational Development, 22*(6), 565–577.

Hildebrand, G. M. (1996). Redefining achievement. In P. F. Murphy & C. V. Gipps (Eds.), *Equity in the classroom* (pp. 149–171). London: Falmer Press.

Hospedales, C. E. (1981). Prestige and success in schools. *Trinidad and Tobago Education Forum, 2*(3), 1–18.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63*(4), 584–601.

Hurtz, G. M., & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the Angoff method? A generalizability theory study. *Educational and Psychological Measurement, 59*(6), 885–897.

Jackson, C. (2003). Motives for 'laddishness' at school: Fear of failure and fear of the 'feminine'. *British Educational Research Journal, 29*(4), 583–598.

Jennings, Z. (2001). Teacher education in selected countries in the Commonwealth Caribbean: The ideal of policy versus the reality of practice. *Comparative Education, 37*(1), 107–134.

Joint Committee on Testing Practices. (2004). C*ode of fair testing practices in education*. Washington, DC: Author. Retrieved, September 6, 2005, from http://www.apa.org/science/FinalCode.pdf.

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

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kane, M. (2004a). Certification testing as an illustration of argument-based validation. *Measurement: Interdisciplinary Research and Perspectives, 2(3), 135–170.*

Kane, M. (2004b). The analysis of interpretive arguments: Some observations inspired by the comments [Rejoinder]. *Measurement: Interdisciplinary Research and Perspectives, 2*(3), 192–200

Kellaghan T., & Greaney, V. (1992). *Using examinations to improve education: A study in fourteen African countries.* Washington DC: World Bank.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioural research.* Washington: APA.

Kutnick, P. (2000). Boys, girls and school achievement: Critical comments on who achieves in schools and under what economic and social conditions achievement takes place: A Caribbean perspective. *International Journal of Education and Development, 20*(1), 65–84.

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*(2–3), 81–101.

Linde, G. (2003). A journey with Durkheim through an examination driven school system. *Educational Studies, 29*(4), 323–335.

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.

MacKenzie, C. G. (1989). The eleven-plus examination in developing countries: A case study. *Educational Studies, 15*(3), 281–300.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A brief introduction to evidence-centered design* (CSE Report 632). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, UCLA. Retrieved, January 14, 2006, from http://www.cse.ucla.edu /reports/r632.pdf.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Murphy, P. (2000). Equity, assessment, and gender. In J. Salisbury & S. Riddell (Eds.), *Gender, policy and educational change: Shifting agenda in the UK and Europe* (pp. 134–152). London: Routledge.

Myhill, D. A. (2005). Testing times: The impact of prior knowledge on written genres produced in examination settings. *Assessment in Education, 12*(3), 289–300.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*(3), 241–286.

Overdorf, J. (2004). Failing the grade. *Far Eastern Economic Review, 167*(26), 54–56.

Parsons, E., Chalkley, B., & Jones, A. (2000). School catchments and pupil movements: A case study in parental choice. *Educational Studies, 26*(1), 33–48.

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

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

Rowan, L., Knobel, M., Bigum, C., & Lankshear, C. (2002). *Boys, literacies and schooling: The dangerous territories of gender-based literacy reform.* Buckingham, UK: Open University Press.

Ryan, J. M., & DeMark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. A. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 67–88). Mahwah, NJ: Lawrence Erlbaum.

Sacks, P. (2001a, June 8). How admissions tests hinder access to graduate and professional schools. *Chronicle of Higher Education, 47*(39), 11.

Sacks, P. (2001b, December/January). Pseudo-meritocracy: A response to "The future of affirmative action" by Susan Sturm & Lani Guinier. *Boston Review*. Retrieved December 26, 2005, from http://bostonreview.net/BR25.6/sacks.html.

Schafer, W. D. (2003). A state perspective on multiple measures in school accountability. *Educational Measurement: Issues and Practice, 22*(2), 27–31.

Schagen, I., & Elliot K. (Eds.). (2004). *But what does it mean? The use of effect sizes in educational research.* Slough, UK: NFER.

Schrag, F. K. (2004). High stakes testing and distributive justice. *Theory and Research in Education, 2*(3), 255–262.

Siegel, H. (2004). High-stakes testing, educational aims and ideals, and responsible assessment. *Theory and Research in Education, 2*(3), 219–233.

Stanley-Marcano, J., & Alexander, M. C. (1998). Trinidad and Tobago [Country report]. In M. Bray & L. Steward (Eds.), *Examination systems in small states: Comparative perspectives on policies, models and operations* (pp. 113–119). London: Commonwealth Secretariat.

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy, and Practice, 12*(3), 275–287.

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. Division of Educational Research & Evaluation. (2002). *Secondary entrance assessment guidelines 2003–2004*. Port of Spain, Trinidad: Author.

Trinidad and Tobago. Division of Educational Research & Evaluation. (2004). *A proposal for a change in weight and modification of the table of specifications of the Secondary Entrance Assessment (SEA) papers* (2005–6). Port of Spain, Trinidad: Author.

Trinidad and Tobago. Task Force for the Removal of the Common Entrance Examination. (1998). *Report.* Port of Spain, Trinidad: Ministry of Education.

Walker, D. A. (2004). The importance of drawing meaningful conclusions from data: A review of the literature with meta-analytic inquiry. *NASPA Journal, 41*(3), 452–469.

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.

Willingham, W. W. (2002). Seeking fair alternatives in construct design. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), The *role of constructs in psychological and educational measurement* (pp. 193–206). Mahwah, NJ: Lawrence Erlbaum.

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

Young, M. J., & Yoon, B. (1998). *Estimating the consistency and accuracy of classifications in a standards-referenced assessment (CSE Technical Report 475).* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing, UCLA. Retrieved, January 26, 2006, from http://www.cse.ucla.edu/Reports/TECH475.pdf.

Zieky, M. (2002). Ensuring the fairness of licensing tests. *CLEAR Exam Review, 12*(1), 20–26.

Zwick, R. (2002). *Fair game?: The use of standardized admissions tests in higher education.* New York: RoutledgeFalmer.

Zwick, R., & Schlemer, L. (2004). SAT validity for linguistic minorities at the University of California, Santa Barbara. *Educational Measurement: Issues and Practice, 23*(1), 6–16.