Caribbean Curriculum
Vol. 15, 2008, 71–114

CAN STANDARDS-REFERENCED, LARGE-SCALE ASSESSMENT DATA LEAD TO IMPROVEMENT IN THE EDUCATION SYSTEM?
Judging the Utility of Student Performance Standards in the Primary School National Assessments of Educational Achievement

Jerome De Lisle

This paper documents the development of performance standards for the Trinidad and Tobago primary school national assessments of educational achievement. Performance standards are written expectations of student achievement operationalized in defensible cutscores. A major argument in this paper is that these standards are necessary to evaluate quality in the education system because they directly address the question of *"how good is good enough?"* Standards-referenced measurement systems are the basis of both compensatory and accountability systems. Standard-setting procedures in the 2005 and 2006 national assessments of educational achievement are described, followed by an evaluation of procedural validity using quantitative and qualitative data gathered from panellists. The findings indicated that while panellists appear confident about procedures and outcomes, cognitive complexity and organizational inefficiency could prove to be critical constraints. While the introduction of student performance standards appears useful; by itself, it cannot lead to significant education improvement unless there is also a coherent policy for effective data use within a national evaluation system. In developing such a policy, consideration must be given to choosing between (a) low versus high examination stakes, (b) compensatory versus accountability policies, and (c) school-based versus centralized management of test processes. At the very least, stakeholders must understand the purpose of national assessments of educational achievement.

> *In the enthusiasm for national testing, there is sometimes a fallacy that measurement by itself will induce positive change in education. Even when advocates recognize that testing must be linked to further action, just what that action should be is often unclear. (*Chapman & Snyder, 2000, p. 458*)*

Examining the Context of Large-Scale Assessments

Eleven Plus Testing as Legacy and Tyranny

History has it that the first public examinations were the rigorous civil service examinations of ancient China (Miyazaki, 1976; Wilbrink, 1997). Just as with high-stakes examinations today, the results were released publicly with either favourable or grave consequences for both individuals and their families. By the 18th century, competitive public examinations were also introduced into a number of schooling systems. Today, while public examinations are common in many schooling systems, the way in which measurement tensions from different assessment purposes are balanced largely depends upon tradition and policy priorities (Broadfoot & Black, 2004). Sadly, in many Caribbean education systems, high-stakes public examinations are the dominant large-scale assessments (Payne & Barker, 1986). In particular, selection at Eleven Plus is, as a corrupting legacy, reducing the emphasis on classroom assessments and diminishing the value of newer instructionally supportive large-scale assessments. Perhaps, this distortion is most evident in the organization and mission of the primary school, with its manic focus on ensuring that students are prepared for Eleven Plus success, at all costs. Indeed, in order to guarantee success, early segregation of students is required to facilitate homogenous classrooms and appropriately paced instruction (Hanushek & Wobmann, 2006). Under these conditions, there is little early diagnosis and intervention. While these arrangements ensure that some high-ability students excel, teaching and organizing to a high-stakes test has negative consequences for most (Evans, 2001).

In the past, without large-scale assessments designed specifically for monitoring learning outcomes, Caribbean education systems lacked the data necessary to judge quality (World Bank, 1993). Therefore,

instituting national assessments of educational achievement became an important pillar in the education reform strategy promoted by international funding bodies (Sebatane, 2000). Ironically, at the same time, in developing countries, proposals for measurement-driven education reform using high-stakes testing were being promoted by some (Eisemon, 1990; Heyneman, 1987). For example, Chapman and Snyder (2000) had argued for the "positive impact" of high-stakes public examinations, citing the work of London (1997) in Trinidad and Tobago. However, London's claim that the Eleven Plus essay component encouraged the teaching of writing in the classroom was based purely on anecdotal data. Kellaghan and Greaney (2004) were more realistic in assessing the use and consequences of high-stakes public examinations. While alluding to the possible benefits of measurement-driven instruction, they readily admitted to the negative consequences. Perhaps, with hindsight, it was extremely naïve to believe that a single examination, high stakes or not, could have a sustainable, positive, and system-wide impact. Such a conclusion underestimates the complexities and unpredictability of the washback phenomena (Broadfoot, 2002; Wall, 2005).

Can a High-Stakes Examination Tell Us About Achievement Standards?

National assessments of education achievement are standardized, large-scale measures designed primarily to describe levels of achievement in parts of or the whole education system (Greaney & Kellaghan, 2007). National assessments originated in the decisions made at the World Conference on Education for All (WCEFA) held in Jomtien, Thailand in March 1990 (Kellaghan & Greaney, 2001). However, by 2000, assessments designed to monitor achievement standards were still relatively rare in the region. Indeed, in 1992, only Jamaica and St. Lucia had developed adequate monitoring systems (World Bank, 1993). In the absence of such measures, some policy makers have resorted to using data from high-stakes public examinations to evaluate achievement standards (Trinidad and Tobago Chamber of Industry and Commerce, 2006; Trinidad and Tobago. Ministry of Education [MOE], 2005). While this approach might appear useful, the implication of such a strategy has not always been adequately gauged. Certainly, the quality of information derived from placement examinations like the Eleven Plus is limited by

content validity, cognitive complexity, reliability, teaching to the test, and student opportunities to learn.

Even if using data from a public examination was appropriate in evaluation, the information from the Eleven Plus is entirely norm-referenced and so cannot answer the question, *"how good is good enough?"* (Brandon, 2005). This question is an important one and relates to the adequacy of student performances. An answer always requires either criterion- or standards-referenced data. Unfortunately, without formal criterion-referenced cutscores, some technocrats have chosen to invent capricious cutscores, such as 30% or 50% of the total score, without reference to the difficulty of the test (MOE, 2005). Indeed, the usefulness of the 30% cutscore on the Eleven Plus as an indicator of "achievement standards" is limited because the working of the placement system depends primarily on ranked scores and not mastery of content. There is no assurance that students below the 30% cutscore are performing poorly; rather they are simply doing worse than the others. Of course, as highlighted earlier, the core problem is misuse of data from high-stakes public examinations for evaluation. National assessment policy, therefore, should distinguish between public examinations for selection and certification, national assessments for monitoring system quality, international assessments for international benchmarking, and classroom assessments for learning and feedback (Braun & Kanjee, 2006; Fiske, 2000).

The increasing role of national assessment systems within developing countries is an example of globalization and international transfer of educational reform strategies from industrialized societies (Benveniste, 2002; World Bank, 1993). Unfortunately, in the Caribbean, this transfer has usually been implicit, contained in the policy documents of the funding agencies. Less often, Caribbean Ministries of Education are active borrowers of assessments. It is rare that local policy makers seek out appropriate strategies and modify protocols to ensure contextual relevance and effectiveness (Sebatane, 2000). This is worrying because implementation fidelity for assessments will be affected by the values a society holds towards tests and test use (Oakland, 1995). Indeed, Greaney and Kellaghan (1996) have noted the many constraints for implementing national assessments within developing countries, including lack of measurement capacity. Ultimately, then, building

institutional capacity is one of the keys to successfully implementing a viable national assessment system. Currently, in Trinidad and Tobago, the Division of Educational Research and Evaluation (DERE), in the Ministry of Education (MOE), holds the primary responsibility for national assessments of educational achievement (T&T. DERE, 2006). Many of the recent improvements in system functioning have been due to better staffing, resourcing, and training coupled with leadership at technical and political levels. However, new issues are emerging to threaten further progress, including lack of stakeholder understanding, the absence of policies to guide evidence-based decision making, and inappropriate or limited data use.

Coming to Grips With the Purpose of National Assessments of Educational Achievement

Despite the implementation difficulties, it seems likely that data from national assessments of educational achievement are essential to improving education in Latin America and the Caribbean. The role of a national assessment system is magnified, perhaps, by the pervasive problem of low quality and high inequality in the region. Notable, too, is the apparent complexity of achievement patterns across geographical regions and social groups (Winkler, 2000). Indeed, the first step towards resolving inequity is the provision of disaggregated data needed to identify the nature and size of achievement gaps (Lewis, 2005). From the standpoint of the major funding agencies, the need for such an efficient national assessment system in Trinidad and Tobago was never in doubt. Therefore, building capacity became an early goal of the reform policies of international funding agencies (World Bank, 1993). World Bank reports also implied that the lack of national monitoring and evaluation systems was to blame for the pervasive inattention to equity and efficiency issues (World Bank, 1995). These concerns fuelled the reforms in the DERE in the late 1990s.

The need for quality data on the performance of the education system should have been obvious to local technocrats and academics on examining findings from the first international assessment in which Trinidad and Tobago participated. The assessment of reading literacy conducted by the International Association for the Evaluation of Educational Achievement (IEA) in 1990 and 1991 showed that Trinidad

and Tobago performed poorly (In the Grade 3 assessment T&T was ranked 25th among 27 other countries, which were mostly members of the OECD.) (Elley, 1992). The data clearly pointed to low quality and pervasive inequality across schools and regions (World Bank, 1995). Sadly, 16 years later, as evidenced in the country's performance in the 2006 Progress in International Reading Literacy Study (PIRLS), equity and quality are still critical issues (Mullis, Martin, Kennedy, & Foy, 2007). Therefore, a functional national monitoring and evaluation system is needed to (a) identify achievement gaps, (b) determine factors associated with these achievement gaps, (c) provide information on improvements resulting from education reform initiatives, and (d) facilitate the development of policies to support underperforming institutions and regions. A measurement system that supports these functions requires quality instruments, appropriate sampling strategies, performance standards, efficient analysis of data, and appropriate policies for data use (Greaney & Kellaghan, 1996).

Prior to 2004, a rudimentary system of "national testing" existed in the form of a centrally designed test distributed to all schools, with teachers expected to score student papers and provide feedback to students. Divisional facilitators were sometimes asked to assist in this process. However, this system could not meet all the objectives of a "real" national assessment system because little information on achievement standards was provided. For example, prior to 2004, only 10% of scripts were centrally scored (Manning, 2004). The system lacked integrity and purpose, and the function of the test remained nebulous. Consequently, in 2004, the MOE chose to install a more efficient, centrally administered national assessment programme. Instruments were designed to measure both achievement and non-achievement variables. The achievement tests in language and mathematics were administered to the entire population of Standard 1 (6- to 7-year-olds) and Standard 3 (9- to 10-year-olds), whereas the measures of teacher and student attitudes were administered to a sample of schools. Compared to the earlier system, there was a greater degree of centralization coupled with radical improvement in the core aspects of test development, such as item writing, scoring, setting performance standards, and data analysis.

Installing a "New" System for Monitoring Achievement Outcomes

Clarifying the Purpose and Intent of the Trinidad and Tobago National Assessments of Educational Achievement

The 2004 national assessments of educational achievement included pen and pencil tests covering four content strands in mathematics and six in language arts, using a combination of discrete and extended constructed-response formats. Centralized test development and scoring facilities were developed and coordinated by the DERE in conjunction with the Division of Curriculum Development (DCD). Curriculum facilitators played a major role at all levels of test development. Initially, there was some ambiguity about the role of the test and the stakes to be attached. As late as 2006, it was reported that the national achievement tests were to be a part of the continuous assessment process. The implication was that scores from the tests might contribute to high-stakes placement decisions at Eleven Plus (Hazel backs down, 2006). Such a strategy would have significantly increased test stakes, with consequent threats to defensibility and integrity. To be fair, however, this ambiguity of purpose was also present prior to 2004, especially within implementation proposals for the Continuous Assessment Programme (CAP). CAP's vague assessment policy implied that classroom, school-based, and national assessments were similar in purpose and function. It may be that the implication of using scores from school-based assessments for high-stakes placement decisions had not been adequately considered (Jones, 2001).

The design choices of the 2004 national achievement tests led to the following decisions: (a) centralization of test development, (b) census annual administration, and (c) low-stakes implementation. There were advantages and disadvantages to each decision. On the one hand, the advantages of greater centralization were that the quality of the examination would be significantly improved and national benchmarks could be established. Additionally, with census administration, school performance information would be available for all institutions, facilitating the development of a comprehensive national monitoring system. On the other hand, the main disadvantage was that teachers would be faced with an externally mandated large-scale test that they might perceive as unnecessary. As in the United Kingdom (UK), teacher

resistance would then become a significant obstacle (Grant, 2006; Mayo, 2005). The perception that national assessments of educational achievement are an external imposition might also restrict ownership and limit data use. Additionally, with a low-stakes examination, some schools might simply opt out or choose to ignore the results, even with mandatory administration (Heubert & Hauser, 1999). Indeed, the tendency of schools to opt out of government-led reforms had occurred in the past, especially in the government-assisted sector (Braithwaite, 1981).

When a Norm-Referenced System is not Good Enough

In some aspects, the norm-referenced reporting system developed for the 2004 national tests was a significant advance over that used in the Eleven Plus. For example, the use of normal curve equivalents (NCEs) like the standard score ensured interpretable data that could be easily processed. The NCE standard score has a mean of 50 and a standard deviation of 21.06. The NCE scale of 1–99 coincides with a percentile rank scale at 1, 50, and 99. The advantage of the NCE scores is that the intervals between scores are equal and scores can be meaningfully combined and averaged. A national mean score of 50 serves as a benchmark or reference point for comparing the performance of regions and institutions. However, a central argument in this paper is that standards-referenced measurement is still required for sound evaluation. In terms of cost-effectiveness, the question then becomes, *How do we justify the additional expertise, training, cost, and time spent in devising and implementing a new system of performance standards?* Any judgement of cost-effectiveness must consider the inability of norm-referenced data to adequately answer the critical question of *"how good is good enough?"* (Brandon, 2005). Norm-referenced data can only provide information on the relative performances of schools and districts, but not on criterion-referenced standards.

Answering the question of "*how good is good enough?"* requires a benchmark for student performance that is tied to the curriculum. A standards-referenced measurement system is able to provide such an answer because it reports information anchored in substantive statements about the levels of student performance (Linn, 2005). Standard-referenced systems make use of cutscores that are absolute but not

capricious. A capricious cutscore, like the 30% threshold used in the Eleven Plus, lacks meaning because it is simply a point on a score scale, without reference to any particular educational criterion. While professional judgements are central to standards-referenced system, these judgements are valid and defensible to the extent that qualified, trained judges are used, appropriate criteria are constructed, and explicit procedures are closely followed during the exercise. Therefore, establishing the defensibility and validity of standards-referenced cutscores becomes the focus of the next part of this paper.

Developing Defensible Student Performance Standards

The fundamental element of any standards-referenced measurement system is the student performance standard (Hansche, 1998). Kane (2001) defined such a student performance standard "as a level of performance described in terms of what examinees at a particular level know and can do" (p. 55). Therefore, performance standards might be regarded as informed expectations of student proficiency, based on the objectives of the content standards and test items. These standards allow a qualitative description of different levels of performance, which might be considered the "normative" aspect of the standard. Additionally, there is also a substantive aspect of a standard, which refers to the written descriptive statements of the expected performance (Haertel & Lorie, 2004). Standards-referenced systems are similar to criterion-referencing systems because they are based upon student skills and knowledge rather than comparative performance. However, standards-referenced systems differ from criterion-referenced systems because the focus is on *expectations of proficiency* rather than on mastery of specific content. Whereas traditional criterion-referenced systems make use of quantity and category indices, such as the percentages correct or the numbers failing and passing on a particular task or class domain, standards-referenced systems provide qualitative descriptions of student performance (Cizek & Bunch, 2007; Nitko, 1980).

Standard setting is the test procedure used to develop a system of student performance standards. In this process, the cutscore for each proficiency level is obtained. These cutscores are at the heart of the process, providing an operational definition for each proficiency level, with a single cutscore at the lower boundary. By definition, individuals who

have attained marks at or above the cutscore would have met the specified standard. Therefore, a system for performance standards should include (a) definitions of each level of proficiency, (b) standard-setting procedures for obtaining the cutscores for each level, and (c) associated cutscores (Brandon, 2005). Performance standards have been used for years in the licensure and certification industry within the United Sates (US) and Canada (Meara, Hambleton, & Sireci, 2001). However, setting standards at this level is relatively easier because there are usually only two levels—pass or fail—for any licensee. By contrast, education tests require multiple proficiency levels, sometimes as many as six. As a result, the process of standard setting in education is more complex and susceptible to error, with some popular methods harder to implement. Most psychometricians now agree that the current set of standard-setting procedures are rational, well documented, scientific approaches to developing reasonable standards for educational performances (Cizek, 1993; Zieky, 1997).

It is important to distinguish between the "examination standards" in public examinations and the performance standards constructed in national assessments. The method of setting standards in both large-scale assessments is quite different, with different conceptions of rigour and defensibility. In public examinations like the CXC or GCE O Levels, the process of setting grade boundaries is organized very differently and makes use of very different protocols. For example, in the setting of grade boundaries in public examinations, there is usually a central management role for the accountable officer in the examination agency (Baird, 2007; Tomlinson, 2002). In one specific procedure described for the UK Assessment and Qualifications Alliance (AQA), an awarding committee of four to eight members scrutinizes the work of judgemental grades only, with the remaining grade boundaries determined by calculation (called arithmetic grades). Professional judgements by members of this awarding committee will be based on the quality of work, the comparability of students' work, and statistical data from the current and preceding years. Each member of the awarding committee independently reviews a different sample of scripts within the grade boundaries initially recommended by the principal examiner (Meyer, 2005).

In standard setting for national assessments of educational achievement, there is no formal role for any single individual equivalent to the accountable officer of an examination board; instead, decisions are made by a large panel of judges. Secondly, the validity and defensibility of the performance standard is closely tied to following a well-known and published protocol. While defensibility is a legal and technical benchmark, there can be no true standard, and different procedures will give vastly different results (Cizek, 1995). Therefore, defensibility emphasizes procedural and technical accuracy and rigour. Cizek (1993) considered a defensible standard to be "the proper following of a prescribed, rational system of rules or procedures resulting in the assignment of a number to differentiate between two or more conceivable states or degrees of performance" (p. 100). Current best practice in most procedures requires (a) many high quality judges, (b) well-written descriptors, and (c) large number of work samples.

Designing the Standard-Setting Plan

To establish a comprehensive standard-setting system for national assessments of educational achievement, five components are required: (1) the administrative arm, consisting of the implementation team and an authority to set policy; (2) a content domain; (3) selection of persons (judges) to make judgements about desired levels of performance; (4) a methodology for collecting judgements and estimating standards; and (5) some means for reporting the results (Reckase, 2000). A standard-setting plan for Trinidad and Tobago would have to fit the social and cultural milieu and take into account possible barriers to implementation. It was hypothesized that possible barriers to implementation might include: (1) lack of familiarity with standard setting in the educational community, (2) lack of expertise by teachers in both standard-setting procedures and the content areas assessed, (3) variability in the outcomes and procedures of current standard-setting methods, and (4) the absence of Item Response Theory (IRT) item analysis. The last meant that some standard-setting procedures, such as item mapping and the bookmark method, were automatically excluded. The standard-setting plan would also need to consider the (a) length and cost of the process, (b) administrative capacity of the MOE, and (c) the number of potential qualified judges in the system (Greaney, 1996).

Table 1. Preliminary Labels and Performance Levels Descriptions for the 2006 Primary School National Achievement Tests of Trinidad and Tobago

DERE Performance Labels	Alternative Performance Labels Used in the Literature	Description of Performance Level
Level 4	Advanced, Exemplary, Accelerated, Distinguished Exceed Standards	**(Exceeds the Overall Standard of Work required at this Level):** Superior academic performance indicating an in-depth understanding and exemplary display of the skills required.
Level 3	Proficient, Competent, Mastery, Satisfactory, Meets Standards	**(Meets the Overall Standard of Work required at this Level):** Satisfactory academic performance indicating a solid understanding and adequate display of the skills required.
Level 2	Partially Proficient, Nearing Proficiency, Nearly Meets Standards, Borderline, Approaching Basic, Just Below Standards	**(Nearly Meets the Standard of Work required at this level):** Marginal academic performance, work approaching, but not yet reaching, satisfactory performance. Performance indicates a partial understanding and limited display of the skills required.
Level 1	Below Basic, Beginner, Novice, Emergent, Minimal, Developing, Well Below Standards	**(Well Below the Standard of Work required at this level):** Inadequate academic performance that indicates little understanding and minimal display of the skills required. There is a major need for additional instructional opportunities, remedial assistance, and/or increased student academic commitment to achieve at the Proficient Level.

An ad-hoc implementation committee was constituted consisting of members of the DERE and the consultant. A proposal was put forward detailing the plan for standard setting. The plan included a framework that suggested four to six performance levels and accompanying labels and descriptions (T&T. DERE, 2005). The administrative arm finally agreed to four performance levels with accompanying descriptions, illustrated in Table 1. While the implementation team believed that the 2004 data pattern might have supported a five-level specification, it was decided to begin with a simplified four-level system that would enable judges and users to easily identify passing (Levels 4 and 3) and failing students (Level 1 and 2). For this paper, the preferred performance level labels are (Level 4) *Exceeds Standards*; (Level 3) *Meets Standards*; (Level 2*) Nearly Meets Standards*; and (Level 1) *Critically Below Standards*.

The next step was to select a set of procedures to ensure defensibility and transparency. The implementation committee considered these two attributes critical to eventual acceptance of the standards by policy makers, teachers, and other key stakeholders. Defensibility, the most important attribute, called for appropriate attention to both substantive and procedural issues. However, choosing between the range of standard-setting methods was a daunting task. Berk (1986) listed as many as 38 different methods and, more recently, Cizek and Bunch (2007) listed 12 popular procedures, each with many variants. The traditional approach has been to classify methods as either test- or examinee-centred (Kane, 1998). Test-centred methods, such as the Angoff and bookmark, are focused on items or the test, while examinee-centred methods, such as the contrasting group and borderline method, are focused on the performance of the candidates. The holistic methods, such as the body of work and booklet classification, are best classified as performance-centred methods rather than test-centred because they make use of student work samples.

Standard-setting methods for national assessments of educational achievement differ in cognitive complexity, reproducibility, precision, and appropriateness (Reckase, 2000). This might be one argument against using a single method, such as the Angoff. While the Angoff has high precision and reproducibility, its higher cognitive complexity might result in difficulty for novice panellists. Indeed, there is debate over the

viability of some procedures used in the Angoff within the education setting (Hambleton et al., 2000; Zieky, 1997). The final decision, therefore, was to use three separate methods and a final synthesis decision. The methods chosen were two variants of the Angoff—whole booklet classification and the contrasting group method. This multiple-method synthesis procedure was originally employed by the Kentucky Department of Education and is further documented in the academic literature by Green, Trimble, Scott, and Lewis (2003).

The DERE synthesis strategy is summarized in Table 2. Each method is identified by the locus and reference of the individual judgement task. As shown, the chosen methods in the triangulated design differ in the nature of panellists' judgement. The Angoff procedure requires judges to estimate the probability of a minimally competent candidate at any achievement level answering the question correctly. There are multiple variants of the method, including some protocols that require judges to estimate the most likely score of students at each level. The latter variant is used for constructed response questions. Whereas the Angoff variants required judgement on expected performances on items, in the whole booklet classification procedure, panellists made a judgement on the entire body of work in the completed test. This post-hoc holistic judgement required judges to classify student performance in one of 12 levels. The variant of this method followed the published procedure in Jaeger and Mills (2001). The 12 levels were also used by panellists in classifying student performances in school on the contrasting group method. Details of each procedure used in 2005 were first published in a workshop manual (De Lisle, 2005) and further adjustments were made in 2006.

Evaluating the Standard-Setting Process

Evaluating the standard-setting process involves gathering evidence for the validity of performance standards. Assuming that there is already sufficient credible evidence for validity of the test instrument, evidence is also needed to support the conceptual basis of the standards and its operationalization into cutscores. There were a number of useful evaluation frameworks in the literature. For example, Schafer (2005), using a sponsor's perspective, proposed an institutional perspective, which subsumed legal, psychometric, and definitional foci.

Table 2. Standard-Setting Strategy for the 2005 and 2006 Primary School National Achievement Tests

Traditional Classification of Procedures	**Test-Centred Methods**		**Examinee-Centred Method**	**Combination**
Standard-Setting Procedures	**Probability & Modified Angoff**	**Whole Booklet**	**Contrasting Group**	**Synthesis**
Classification Based on Nature of Decision-Making Process (Locus & Procedure)/Characteristics	***Binary/Probability Judgements & Estimated Proficiency on Item***	***Binary Judgement & Proficiency on Set of Items***	***Binary Judgement & Proficiency in all School Settings***	***Impact Data and Comparison of Cutscores***
1. Iterative process	Yes	No	No	No
2. Review of student performance records in classes	No	No	Yes	No
3. Examined examination scripts	Round 2	No	No	No
4. Examined normative data, including relative difficulty of items	Yes	Not required	No	No
5. Knowledge of minimum pass levels for each item or cutscores set by other teachers	Yes	No	No	Yes
6. Made use of relative impact data[º]	No	No	No	Yes

[º] Provided in Synthesis only

Schafer successfully applied the 20 evaluation criteria developed by Hambleton (2001) to this institutional perspective. Haertel and Lorie (2004) developed five validity arguments covering content standards, alignment, accuracy and precision, performance standards, and the cutscore. Gomez and Padilla (2004) offered a sixth possible validity argument for the consequences associated with implementing performance standards. Earlier, Kane (1994) had developed a comprehensive framework for validating performance standards, based on three distinct areas: procedural, internal, and external. Hambleton and Pitoniak (2006) only recently summarized each component area, with procedural validity involving explicitness, practicality, implementation, panellist feedback, and documentation.

Significantly, all evaluation frameworks consider procedural validity as critical to legal and technical defensibility and a defence to questions that might be raised about the arbitrariness of the standards. US courts of law consider "arbitrariness" in standard setting as unreasonable and capricious decision making (Carson, 2001). Thus, this evaluation study is restricted to those validity arguments focused on procedural propriety. Traditionally, this information is obtained from panellists' feedback using quantitative methods. For example, Hambleton (2001) developed a standard-setting evaluation questionnaire to measure the adequacy of training, activities, confidence in the process, and outcomes. This evaluation questionnaire includes a mixture of closed and open-ended items and provides primarily retrospective, recognition information when administered at the end of the event (Skorupski & Hambleton, 2005). While useful, this type of data provides only weak support for procedural validity and gives little insight into the cognitive processes of panellists (McGinty, 2005). It is therefore prudent to also collect qualitative data that provides insight into panellists' thinking when making judgements (Skorupski & Hambleton).

Qualitative evaluation data collected concurrently with the judging process will provide greater insight into the actual cognitive decision-making processes of judges (Giraud, Impara, & Plake, 2005; Hurtz & Auerbach, 2003). This type of data will be useful in evaluations that are focused on improving the programmes being evaluated and that are intended to provide in-depth contextualized information on practices (Greene, 2007). Since judges are the key to improving the quality of the

process, the panellists' perspective is a critical source of information in monitoring. There is certainly need for more information on how judges make decisions in standards setting. For example, Brandon (2004) noted that, despite its popularity, the Angoff was a virtual black box when it comes to understanding how panellists make judgements. Plake (2008) also highlighted the need to understand the importance of the orientation and training components. These arguments hold doubly true for standard setting in Trinidad and Tobago, with the lack of standard-setting experience among judges.

Evaluation Methodology

The evaluation study employed multiple methods, gathering different types of data using a questionnaire and an unstructured diary/journal. This approach was designed to provide both concurrent recall and retrospective recognition data. The diary/journal was especially valuable as a data collection tool because it allowed in-depth reflection of participants and an exploration of evolving insight over the duration of the process and the event (Hiemstra, 2001). Panellists were expected to keep an account of their daily experiences as well as to recount their personal understandings, thoughts, and feelings about the process. In 2005, 49 questionnaires and 38 journals were collected. In 2006, 65 questionnaires and 51 diaries/journals were returned. Since the standard-setting procedure was an intensive one-week process, the 2006 return rate of 92% for the questionnaires and 69% for the journals and diaries can be considered adequate.

The questionnaire had a slightly different design for the two administrations. In 2006 only, items from Part 1 of the questionnaire were based entirely on the evaluation framework of Hambleton (2001). The instrument included closed and open-ended questions targeting retrospective recognition of key standard-setting activities, such as construction of descriptors, efficacy of training, and overall confidence in the outcomes. Part 2 of the 2005 and 2006 questionnaire consisted of eight open-ended questions originally designed by the consultant for use in a focus group interview. In 2005, the questions were:

1. Did you understand the standard-setting process?
2. What did you like or didn't like about the process?

3. What were your personal understandings of the performance levels and did they match what you were required to do?
4. Which standard-setting methods were most difficult in terms of understanding what was required and what you were required to do?
5. Which standard-setting method in your mind best captures the levels of student performance?
6. To what extent do you agree with the process and the standards derived?
7. Did you benefit from the discussion within your group and how did that discussion change your individual ratings?

In 2006, a number of additional questions were added, including:

1. Explain the strategy you used to assign students to the different categories.
2. Were there any problems that might have influenced the way you categorized students?
3. To what extent do you agree with the process and the standards derived?

All descriptive quantitative data were collated and presented in a tabular format. Qualitative data from the two years were transcribed, coded, organized, and reviewed by two independent researchers. Data analysis focused on exploring the themes that emerged from the data rather than applying a coding scheme based on Hambleton and Pitoniak's (2006) specifications for procedural validity. Content analysis was first used to identify the recurring themes, which were then coded. This preliminary coding was constructed after an initial reading of the text. Themes and sub-themes were then re-grouped and organized into tables. The final coding sets of both reviewers were then reconciled. The reconciled tables listed themes, codes, and sample statements constructed. The entire coding scheme was then independently reapplied to the text data and the tables further refined. Further reconciliation was attempted before constructing the final sample statements, codes, and themes. Following the finalization of the themes and codes, the narrative was written by decoding each of the tables.

Evidence for Procedural Validity

Findings From the Questionnaire

Table 3 provides self-report data from the 2006 panellists on their perceived success on different standard-setting tasks. As shown, the majority of panellists reported success on all seven standard-setting tasks. The highest rate of perceived success (successful and very successful) was on completing the test (98.4%), but judges also felt successful on the integrated training and practice exercises. However, 22.2% of the panellists felt they had achieved only partial success setting the descriptors. Likewise, for the Angoff procedure, 17% of the panellists reported only partial success on the Round 1 Angoff task. This figure, however, dropped to 10.9% in Round 2 and 4.7% in Round 3. This suggests that the ability to make judgements improved in later rounds, possibly due to greater opportunities for practice. Table 4 shows panellists' perception of the performance standards produced. Overall, panellists felt that the quality of the standards generated was adequate, with 96% giving an adequacy rating of 3 to 5 for Level 4; 98.2% for Level 3; 98.2 for Level 2; and 96.6% for Level 4. It appeared, then, that panellists were more confident about the integrity of performance levels at the extreme ends of the scale, including Level 4 (40.8%) and Level 1 (43.1%). However, they were comparatively less confident about the quality of the standards in the middle categories, Levels 2 (31.6%) and 3 (28.1%).

Table 5 displays data on panellists' confidence in the outcomes and the different standard-setting methods used. Supporting the data presented in Table 4, panellists were most confident about the standards in Level 4 and 1, with 36.9% indicating that they had high confidence in Level 4 and 43.1% in Level 1. However, 21.5% of the panellists expressed a moderate level of confidence in Level 3 and 24.6% in Level 2. When responding to the close-ended items, panellists considered the contrasting group and whole booklet classification methods to be the best procedures for setting cutscores, with only 14.5% of the respondents answering in the affirmative for the Angoff. Table 6 listed seven possible factors that influenced panellists' judgements. The factors were ranked in order of importance, with descriptors (65.1%), classroom experience (58.5%), and student responses (53.8%) seen as most critical. The least important

factors appear to be first round judgements and judgements made from the Angoff.

Table 3. 2006 Panellists' Perceptions of Success on Different Standard-Setting Tasks (N = 60–65)

Standard-Setting Activities	Not Successful	Partially Successful	Successful	Very Successful
Completing the Test	0	1.6	41.3	57.1
Angoff Round 3	0	4.7	57.8	37.5
Developing Descriptors	0	22.2	49.2	28.6
Angoff Round 2	0	10.9	62.5	26.6
Angoff Round 1	0	17.2	65.6	17.2
Contrasting Group	0	8.1	75.7	16.2
Integrated Training and Practice	0	8.3	83.3	8.3

Table 4. 2006 Panellists' Perceptions of Quality in the Performance Standards Generated (N = 49–58)

Performance Level	Totally Inadequate		Ratings		Totally Adequate
	1	2	3	4	5
Level 4 (Mastery)	2.0	2.0	12.2	42.9	40.8
Level 3 (Proficient)	1.8	0	17.5	56.2	28.1
Level 2 (Borderline Proficient)	1.8	0	19.3	47.4	31.6
Level 1 (Remedial)	3.4	0	8.6	44.8	43.1

Table 5. 2006 Panellists' Confidence in and Comparison of Standard-Setting Methods, Outcomes, and Procedures (N = 65)

	Confidence			
	Very High	**High**	**Medium**	**Low**
Level 4 Classification	36.9	53.8	9.2	0
Level 3 Classification	16.9	61.5	21.5	0
Level 2 Classification	20.0	55.4	24.6	0
Level 1 Classification	43.1	47.7	9.2	0
Overall Procedure	12.3	69.2	15.4	0
Method	**Percentages indicating particular method as "best" procedure**			
Angoff	14.5			
Whole Booklet	41.8			
Contrasting Group	43.6			

Table 6. 2006 Factors Used by Panellists in Classifying Student Performances (N = 65)

Factors	**Not Important**	**Somewhat Important**	**Important**	**Very Important**
Descriptors at each level	0	3.2	31.7	65.1
Own classroom experience	1.5	7.7	32.3	58.5
Student responses	1.5	4.6	40.0	53.8
Perception of test difficulty	3.1	7.7	40.0	49.2
Discussion in group	0	20.0	36.9	43.1
First round judgements	11.9	33.9	33.9	20.3
Judgements from the Angoff	6.3	25.0	51.6	17.2

The Qualitative Enquiry

The qualitative data included retrospective recall and recognition information from the open-ended part of the questionnaires and concurrent recall information from the five-day journal. Analysis of the text data established five main themes in the respondents' comments. In order of frequency, these were: (1) outcomes related to the performance standards; (2) organization of the process, including group discussion; (3) levels of the performance standards; (4) cognitive challenge of the process; and (5) advantages and challenges of specific standard-setting methods. Focusing first upon outcomes, panellists often reflected on the value and use of the standards for the education community, as the following comment by one panellist illustrates:

This process clearly demonstrates or assists me in understanding how important our role as stakeholders are in terms of coming together with dedication and commitment to formulating standards that would serve the needs of all children across Trinidad and Tobago. [Panellist-Diary]

Many teacher panellists felt that the standard-setting process helped them gain deeper insight into student performances within the classroom. These teachers considered that they might be used to modify expectations of students. For example, one teacher noted that performance standards were useful as benchmarks within her classroom:

The process, when completed, painted a clear picture about the students' performance. After making judgments, it was easy to see the level of our children and to be well equipped to make informed decisions about going back to the classroom and putting my house in order. [Teacher Panellist 1]

An often-repeated theme was the reality check experienced when panellists confronted students' actual work performances in the test. This revelation occurred after panellists had examined students' scripts during the whole booklet classification method or after being presented with test statistics during Round 2 of the Angoff. Most judges were surprised by the poor quality of student responses in the scripts and low mean performance on some test items. This reality check was a critical part in the decision-making process, often leading panellists to lower their initial

expectations. More significantly, this reality check resulted in some panellists re-evaluating the quality of teaching and learning within the nation's schools; a notable point revealed in the following 2006 journal entries of two panellists:

We completed the 3rd round of the Angoff today, using the statistics from the test. It was amazing that the results were so low when compared. There were also a lot of children getting zero in items that should be simple for them. [Panellist 1]

Today, the groups were involved in marking the sample scripts using the whole-book method. We marked STD3 in 4 rounds and Std 1 in 5 rounds. There were not many students found in the upper ranges of the level 4 but quite a few in the lower ranges of level 1. This is good information to be used in better understanding the way children think and learn. . . . There are many children at risk and we must act now. [Panellist 2]

However, to be sure, not all comments about outcomes of the standard-setting process were positive. For example, some panellists felt that skills identified in the standard neither fully matched nor extrapolated to the outcome domain. Additionally, some panellists believed that none of the existing procedures captured fully the range, variety, and complexity of student performance in schools and classrooms. These panellists often argued that, in reality, there were too many factors influencing student competence. While these concerns were often based upon the narrowness of the test and limitations of content standards, some of these panellists were also concerned with limitations in the standard-setting procedures. Indeed, quite a few of this group felt that there was a mismatch between expectations in the standard and real-world student performances, as revealed in the following questionnaire response:

The process was done without any information of school, teacher, or pupil. There are many [other] variables to consider [such as]: (1) completion of syllabus, (2) mental disposition of students on a one-day test, (3) socioeconomic and psychosocial issues affecting children, and (4) children's ages and promotion by attainment. [Panellist-Questionnaire]

In the 2005 standard-setting process, a number of panellists commented negatively on the overall organization of the event. This perceived disorganization was discomforting, leading to frustration and confusion during the judgement process. While administrative delays and alterations might have been relatively infrequent, unsatisfactory arrangements were especially burdensome because they made a difficult and challenging task even more arduous, impacting upon the quality of judgements. For example, one panellist noted that having groups and individuals shifted about constantly added to the confusion over the process:

I did not like the way the group was shifted around constantly with no clear indication of what was to be done. This phase was a bit confusing. However, I understand that since this is the first time that this exercise is being done, adjustments were being made as we went along. [Panellist, 2005]

The situation was much improved in 2006, with better accommodation, quality meals, and additional administrative staff. Thus, in the second year of the process, panellists were more likely to commend the MOE for the physical setting and better organization. For example, one panellist expressed her feelings about the new venue for the workshop in 2006:

This morning I set forth for ____ auditorium in _____ for the standard setting of the National Tests 2006. As I entered I was amazed at what befell my eyes, the setting was pleasing to the eye. . . . This was the first time I was attending a workshop/session held by the Ministry of Education in such style. What a difference! Finally, teachers are being treated as the professionals they are! [Panellist, 2006]

In this exercise, group discussion at the end of each Angoff round appeared to be a critical component of the judgement process, ensuring that different ideas about student performance were disseminated and unrealistic or biased expectations moderated. Without prompting, panellists often noted this aspect in their journals, highlighting the value of group discussion in generating deeper insight into the process of judging. Significantly, in the 2006 standard-setting exercise, a number of panellists also indicated that the group process enhanced their sense of belonging, which made the overall task easier by ensuring quicker social

adjustment to the unfamiliar surroundings. Asked specifically about the role and value of group discussion, panellists indicated five main advantages provided by these discussons: (1) useful information for altering initial expectations and judgements ratings; (2) deeper, thought-provoking insight into student performance for panellists who were new teachers; (3) credible arguments for cementing and making judgements; (4) added insight into students' thinking when responding to written questions; and (5) alternative perspectives on issues and responses. The perceived value of group discussions in the decision-making process is clearly noted in the statements of the following two panellists:

The discussions were extremely healthy. Some people had strong views, but provided credible arguments to substantiate those views. In some instances, it caused me to change my ratings wile in others it didn't since I am also one who would argue strongly for anything that I truly believe in. [Panellist 1]

Group discussions, helped me to expand my way of thinking about students at a Standard 1 level, since my beliefs were based entirely upon interaction with students in my class. The group discussions shed light on very dark matters, so, yes I did benefit from these discussions. The number of years experience in the teaching profession placed me at a slight disadvantage from the other seasoned teachers in my group. I was therefore, very open and willing to accept any thoughts or ideas that they were willing to share. [Panellist 2]

Panellists often commented at length on the nature and role of performance standards, highlighting seven main areas: (1) the normative aspect, (2) the substantive aspect; (3) the measurement instrument; (4) the standard-setting methods employed; (5) differences in their own understanding and expectations; (6) the need for challenging expectations; and (7) outcomes, such as the workability of standards and procedures. On recall, panellists most often emphasized either the substantive element, such as the value of the criterion-referenced descriptors, or the normative element, such as the nature of differences among students at the various proficiency levels. Generally, panellists felt that the real value of the system was the ability to judge student performance using criterion-referenced statements. They also believed that teachers within their classrooms could directly apply both the

substantive and normative elements. This application might prove critical to improving learning outcomes, as indicated by the following two panellists' comments:

I liked the formulation of the descriptors for the various content strands. These descriptors gave a clear indication what one is looking for when doing the analysis of the test items. However, what I did not like, or should I say [was] uncomfortable with, was that I felt my judgments might be too subjective. [Panellist 1]

The process, when completed painted a clear picture about the students' performance. After making judgments, it was easy to see the level of our children and to be well equipped to make informed decisions about going back to the classroom and 'putting my house in order'. [Panellist 2]

Although most teachers placed a high value on establishing challenging expectations for student performance, some panellists did question the measurement process, including the quality of the test and even the particular standard-setting methods employed. To some extent, this reflected a mature view of the assessment process, since deficiencies in the test and measurement process would invalidate the performance standards, as one panellist observed in her journal:

The [whole booklet classification] method does not seem to be a fair process as the results are based on your perception [rather than] taking into consideration the child. The Angoff method uses statistics and the descriptor also does not focus on the child, as there are the exceptions in both cases. The standard setting process cannot be based on one test. [Panellist-Journal Entry]

It was expected that the standard-setting process would be cognitively challenging for most panellists. Although many possessed at least a first degree in a content area or education, few had prior knowledge or experience with standard setting. Therefore, most of the content and methodology were new and unfamiliar to participants. This added greatly to the challenge, as one experience test scorer noted in an entry on the first day of her journal entry:

Today was quite interesting and confusing. Many concepts introduced were quite new. It seemed like quite a task to complete. I am quite interested in understanding the standard setting as the Angoff, [Whole Booklet Classification] and the Contrasting Group. This exercise interested me because after correcting the National Test scripts, I was eager to learn how it was going to be standardized. The process of determining descriptors for the various levels was a tedious one, since agreement was hard to reach at times. However, it was a very good method of knowing the descriptors. . . . At the end of the day I felt I did not have a clue about the Angoff method and I hoped that the next day would assist with this. [Panellist-Journal Entry]

The exercise was also regarded as especially difficult because respondents had to participate in multiple tasks and use information from multiple sources, all at the same time. For example, in making judgements about student performance, panellists were required to construct and use descriptors, examine data, observe student work samples, while discussing their own judgements with those of other panellists. Some processes also demanded skills that panellists may not have acquired. To illustrate, one panellist commented on the difficulty of dealing with the item analysis data presented in Round 2 of the Angoff:

The afternoon session, with the introduction of the statistical evidence, in terms of percentages, brought an even deeper level of reasoning for me. I now had to look at the percentages and determine the difficulty level of the item, look at the rubric and the item itself and determine my round 3 cutscores for the standard 1 Maths. This was a very draining exercise. I do not believe that thinking about anything over my holidays so far drained me that much. I just hope that I performed the exercise well. I dread to think of doing it again tomorrow for Standard 3 Maths. [Panellist-Journal Entry]

The degree of challenge was also high because judges had to learn and employ three different procedures at roughly the same time. While journal entries indicated that panellists' knowledge evolved rapidly, this sharp learning curve sometimes resulted in frustration. The integrated design of the exercise also made it difficult to manage time and some respondents believed they had insufficient training and practice.

When asked directly which standard-setting method was most difficult, in the 2006 questionnaire, 39 of 53 panellists identified the Angoff. The Angoff was considered especially difficult because it was abstract, required interpretation of statistical data, lacked authenticity, and employed the concept of a borderline or minimally competent student, which was difficult for some to conceptualize. Only seven panellists regarded the whole booklet classification method as the most difficult. These panellists gave a variety of reasons such as tediousness, lack of fairness, and difficulty in making a compensatory, holistic judgement. However, when asked directly which method best captured student performance levels, in 2005, 18 out of 41 said the Angoff, and in 2006, 30 out of 56. This compared with 23 out of 45 for the booklet classification method in 2005 and 26 out of 56 in 2006. Thus, surprisingly, although the majority of judges regarded the Angoff as the most difficult, it was also the most highly valued procedure. Panellists who favoured the Angoff pointed to the fact that it was item-based and made consistent use of the descriptors. Many panellists therefore considered this approach as more "objective" because their professional judgement was focused on expectations of performance for each item. Panellists who favoured the whole booklet classification method felt that (a) it better accommodated the multiple factors that could influence student achievement, and (b) was more authentic because it was based on real students' responses.

An analysis of the journal comments revealed some additional reasons for panellists placing high value on the Angoff, despite its cognitive complexity. It seemed that in many cases, panellists regarded judgements on the whole booklet classification method as simply too subjective. Some panellists also reported that it was very difficult to find scripts within a folder that matched the entire ranges of performances; although the scripts in every folder were always graded by percentile scores. In addition, having to make judgements across 12 different performance levels often proved difficult, as noted in the following comment:

The [whole booklet classification] method for standard setting seems to me to be more difficult to assess. The range between the descriptors is hard to decipher. For example what makes the difference between a LEVEL 1-M and a LEVEL 1-H? I did the exercise of evaluating the test

scripts (Whole Booklet Method) but I wasn't comfortable with it. [Panellist-Questionnaire]

Another issue that contributed to the dislike of the whole booklet classification method was the difficulty in making compensatory holistic judgements, a problem that occurred especially in the Language Arts paper. This was because the tasks in this paper varied greatly in the nature of the assessment tasks and length of responses. Thus, differences in performance were not easily resolved, with some students writing excellent essays but doing poorly on some shorter prompts. For example, in a journal entry, one panellist explained her difficulty in making compensatory judgements:

The [whole booklet classification] method is too subjective. . . .in any particular booklet the student showed strength is one area and weaknesses in others and this had to be weighed [against each other in order] to make a good judgment. For e.g., a pupil [might solve] high order questions in problem solving, but showing no performance with basic number or measurement etc. This may lead to bias where students may make the same total score but because of where they showed their strength they are rated according to certain considerations. [Panellist-Journal Entry]

Discussion

This study documented and evaluated the design and reporting of performance standards for national assessments within the primary school system of Trinidad and Tobago. It was argued that one advantage of standards-referenced measurement over capricious or norm-referenced systems was the ability to provide a direct answer to the question of *"how good is good enough?"* This was achieved by the reporting of performance standards, which categorize student achievement into multiple levels based on the judgements of informed panellists. Information on satisfactory and unsatisfactory performances provided in the standards allows a judgement in terms of whether those standards were met across schools and educational districts. This kind of information is necessary for identifying and resolving patterns of inequality. Identification of achievement gaps is the first step in developing a system policy that ensures that all students have equal

opportunities to learn. Such interventions are especially important in Latin America and the Caribbean because inequality remains a pervasive problem despite the expansion in schooling (Perry, Arias, Lopez, Maloney, & Serven, 2006). Perhaps this issue has been neglected in the past because of the lack of high-quality monitoring data.

The quantitative data revealed that the majority of panellists were very confident about the performance standards constructed. They believed that they had successfully performed most of the key tasks. As to the quality of the standards, while respondents were generally positive, some believed that standards at the extreme levels (mastery and remedial) were superior to the levels in the middle (proficient and borderline proficient). During the process of making judgements, panellists relied on a variety of factors, but were mostly guided by written descriptors and their own classroom experiences. This suggests that an important qualification for judges should be the length and variety of experience in the classroom. Data from the open-ended instrument and the journals suggested that some panellists were negatively affected by the limited time for training. Others were concerned about the organization of the process and the increase in cognitive challenge presented. These are significant issues, which impinge on procedural validity and call for better management and organization coupled with increased training in the future. With hindsight, an integrated design, although cost saving, might not be the best approach in this context, considering that most panellists had little standard-setting experience.

As expected, no single method perfectly captured all the nuances of the standard. Thus, the synthesis procedure was an important aspect of the design. However, the DERE might now want to experiment with alternative approaches such as item mapping, the bookmark, or the analytical judgement method (Cizek & Bunch, 2007). Even if a new method is implemented, group discussion must be retained, as it appears critical to the quality of the decision-making process (Hurtz & Auerbach, 2003). This study adds to the international debate over the utility of different methods of standard setting (Hambleton et al., 2000 Plake & Impara, 2001; Zieky, 1997). While most respondents regarded the Angoff as the most difficult method, they also believed that it produced "better" cutscores. One reason may have been that they perceived the Angoff to be less subjective than the whole booklet classification and

contrasting group methods. It might be that panellists did not value methods that relied solely upon human judgement of work samples, or they might have regarded the broad compensatory judgements as unreliable, difficult, or impossible. Improving the process might be a costly exercise, valuable only if costs can be recovered through effective data use.

Of course, improving standard setting must never come at the expense of improving the quality of the test itself (Downing & Haladyna, 2006). Improving the test requires research into test development, with ongoing validation and alignment studies (Bhola, Impara, & Buckendahl, 2003; Haladyna, 2006). It was of concern, therefore, that a few panellists expressed concern over some items and aspects of test scoring. Technical legitimacy of the test is the foundation upon which all other processes and policies are built. Therefore, performance standards can only be valid if the entire assessment system provides scores that allow useful and meaningful inferences. This does not mean, however, that implementing performance standards now is an entirely futile exercise because validity is not an all or nothing characteristic. What is required is a system of continuous improvement at all test development stages (Greaney, 1996). Perhaps, then, the greatest weakness of the current system is that while there is national testing, a national educational quality evaluation system has not been developed (Gvirtz & Larripa, 2004; Olivares, 1996; Wolff, 2004). Thus, to date, the focus has been solely on collecting and reporting data, and not on using this information to diagnose weaknesses, evaluate education reforms, and create systemic changes. This uninformed use of data increases the potential for error and misuse of the tests. Indeed, policy can be constructed that allows a test to become a weapon in pursuit of some vague political, technocratic, or social agenda (Crespo, Soares, & de Mello e Souza, 2000; Heubert & Hauser, 1999).

The Way Forward: Towards Better Use of the Data

Little can be done in the absence of an overall policy on assessment and evaluation. This requires that the MOE constitute a broad-based technical committee to develop a framework for assessment and data use within a national educational quality evaluation system (Greaney & Kellaghan, 1996). This policy might specify exactly the role and purpose of different

assessments, specifying ways in which data are to be used. The policy should also specify structures, resources, and training to be put in place, including teacher training and public education programmes designed to improve assessment and data for all stakeholders. Turning towards the role of national assessments of education achievement, the critical question that emerges is, "*What policies are to be put in place to make best use of national test data?"* The answer to this question lies in examining current practice within Latin America and becoming an active borrower of international assessment strategy (Sebatane, 2000). The evaluation systems of Latin America provide the best guide for Caribbean national assessment systems because these countries share similar education problems, including the plague of persistent inequality (Ferrer, 2006; Winkler, 2000). Interestingly, Chile's system is older than that of the US (Greaney & Kellaghan, 2007).

Still, there is also much to be learnt about appropriate data use from accountability systems in the more developed countries, such as the US and the UK (Herman & Haertel, 2005). On the positive side, one lesson is that while data might be used to ensure that schools and teachers are held accountable for student learning, publicly disseminated rankings of schools based on weak or inappropriate criteria will do more damage than good (Heubert & Hauser, 1999). Therefore, there are important choices to be made between competing agendas when developing data use policy. The most notable choices relate to decisions about the nature of the test, use of data, and administration of national assessments. The first decision relates to nature of the test, with the choice between high stakes, summative or low stakes, formative. The second decision relates to policy, with a choice between accountability and compensatory emphases. The third decision relates to authority and control, with the choice between centralization versus decentralized data use. These three decisions relate to the core function of national assessments, which is to bring to the notice of policy makers and the public the need for more effective education, to justify the reallocation of discretionary resources with more efficient resource allocation, and to ensure management efficiency (Kellaghan & Greaney, 2004).

Test Nature: High Stakes, Summative vs. Low Stakes, Formative

Some consequences are always attached to any high-stakes test, and, in theory, this might be a reward or sanction. In contrast, a low-stakes test is primarily used to provide information on the system (McDonnell, 2005). However, a national assessment system might also be considered low stakes if the information is used primarily to help low-performing schools improve (Ravela, 2005). Admittedly, there is tremendous pressure on administrators to increase the stakes associated with national assessments (Heubert & Hauser, 1999). However, it is more than likely that a high stakes national assessment will impact negatively upon schools or students. In the case of students, these negative consequences might include grade retention or placement in remedial tracks, whereas for schools and teachers, embarrassment or sanctions might be the possible outcomes. An increase in test stakes will also occur if the focus is to ensure that teachers teach certain topics. Thus, while washback can be positive, the evidence suggests that the impact of any high-stakes assessment is more likely to be negative (Broadfoot, 2002). There is the likelihood, even in the case of national assessments, that teachers will simply teach to the test (Firestone, Schorr, & Monfils, 2004).

It is possible, then, that the associated prestige of reporting high-achieving students in a high-stakes system might lead some schools to neglect low-achieving students. In the primary school, this will impede organizational arrangements designed to ensure early remediation. Nevertheless, a low-stakes programme by itself will not lead to improved student learning instead, unless the policy encourages data use for formative purposes. This approach implies a focus on assisting the teaching and learning process. There are numerous examples of such practice within the Latin American region (Ferrer, 2006). For example, Ravela (2005) documented the low-stakes, formative approach in Uruguay, where results from national achievement tests are used to inform and direct in-service teacher training programmes linked to performance in the examinations and the overall social context.

Use of Data Policy: Accountability vs. Compensatory

The formative approach to national assessment data use is evident in compensatory education policies, which are designed to help

underperforming schools succeed. In Latin America, best practice is found in Chile, Uruguay, and Mexico. There is a dire need for such an approach throughout the region because inequality remains a significant educational and social issue. Contrary to what some might think, educational inequality is also a critical issue in Trinidad and Tobago (Perry et al, 2006; World Bank, 1995). Not only is the link between education and poverty notable, but if the most disadvantaged students cannot receive a high-quality education, that link becomes unbreakable (Perry et al.). Indeed, without quality education, disadvantaged students cannot escape the poverty that characterizes many disadvantaged communities (Shapiro & Trevino, 2004). A compensatory education policy provides an efficient mechanism for directing scarce resources to schools and communities in need of help. Winkler (2000) has classified the different compensatory systems in Latin America using type of intervention and targeting mechanism. Interventions may be either supply or demand and targeting mechanism may be geographic, group, or self.

In Chile, the P900 is a supply, group targeting education policy that facilitates the direct link between national achievement test scores and compensatory funding in the form of materials and technical assistance channelled to low-performing schools. Data from national assessments are one of the starting points of a system diagnosis leading to a comprehensive strategy for improving the performance of low-achieving schools. Notable interventions include teacher professional development, targeted support for students at-risk, pedagogic counselling and guidance, and distribution of educational materials. Mexico also has a supply side geographic targeting compensatory education policy called the Program to Abate Educational Lag [PARE] (*Programa Para Abatir el Rezago Educativo*) operated by the National Council of Education Promotion [CONAFE] (*Consejo Nacional de Fomento Educativo*). The programme is designed to provide extra resources to schools that enrol disadvantaged students (Shapiro & Trevino, 2004). Compared to the P900 programme in Chile, this policy places a greater focus on curriculum materials and local decision making, which might increase the cost-effectiveness and impact of the programme.

Locally, it is common for schools to blame low student performance on factors outside the control of the school, with the pretence that

institutions can do little to change what the home has constructed. Thus, one might argue that in Trinidad and Tobago any compensatory policy must work hand in hand with increased accountability. To be sure, systemic results-driven accountability systems have become a worldwide trend (Anderson, 2005). These are built upon objectives, assessments, instructions, resources, and rewards or sanctions, with the general premise that educators are held accountable for student learning (McDonnell, 2005). In theory, a workable accountability system will be structured to transform schools, teachers, and classroom environments, and under such a policy, student learning failures will be attributed to weaknesses in programmes and practices. Ultimately, though, even under a compensatory policy, school personnel must understand that they are *professionally accountable* for student learning. Such accountability, then, becomes a moral endeavour rather than a legal or policy requirement; it is a philosophy of action captured explicitly in the mission and goals of empowered schools and districts.

Data use in compensatory and accountability systems represents a significant conflict in philosophical and political values (Benveniste, 2002; McDonnell, 2005). Compared to compensatory policies, an accountability system, with public reporting of results and sanctions, will increase the stakes associated with national assessments. As in Chile, an accountability system might also work hand in hand with quasi-market school systems to encourage competition and choice between schools (Benveniste). Certainly, public ranking of schools, as already practised by some local educational divisions, might lead to stakeholders demanding impossible quick fixes. In turn, this might make some schools, especially those in exceptionally challenging and difficult circumstances, much more vulnerable (Muijs, Harris, Chapman, Stoll, & Russ 2004). These are rather unpleasant side effects, which suggest that the best approach at this time might be to emphasize the formative use of assessment data. Such an approach synchronizes with the State's current commitment to poverty reduction and enhanced educational opportunity.

Administration: Centralized Control vs. Empowering Schools and Teachers

In Trinidad and Tobago, national assessments are currently administered by the DERE. As discussed by Greaney & Kellaghan (1996), there are

disadvantages to employing an agency within the MOE to manage a national assessment system. One of these disadvantages within a plural society such as Trinidad and Tobago is the difficulty in fully disaggregating data. Failure to disaggregate data fully means that some achievement gaps remain unidentified. Perhaps a more significant problem, however, is the persistent and dysfunctional focus of school communities on the MOE, pointing to a general lack of ownership and powerlessness. Harvey (1981) referred to such a pervasive sense of powerlessness and argued that it might be exemplified in the tendency to personify and scapegoat the "Ministry" for all the system's ills. These weaknesses may be attributed to the governance structure for education in Trinidad and Tobago and the tendency to centralize key educational processes, including assessment and evaluation. Existing arrangements for national assessments only serve to confirm and validate the role and importance of the centralized structure.

Therefore, such a strongly centralized model might lead to further powerlessness and lack of ownership, further limiting effective data use by schools and teachers. It is notable that in Mexico, the PARE compensatory programme includes a separate school-based management policy called Support to School Management (Apoyo a la Gestión Escolar) [AGE]. This programme includes monetary support and training to parent associations, with consequent involvement of parents in decision making relating to school processes. In the local context, the problem-solving and intervention capacity of school personnel must be enhanced, allowing them to use data to solve site issues of academic underachievement. This approach might include strategies to improve assessment literacy among administrators and teachers. Enacted policy should also provide training and resources for contextualized decision making and problem solving (Earl & Katz, 2006). Such approaches will lead to the empowerment of schools and communities and will lessen the focus on "the Ministry" as the sole source and solution of every local problem.

The current pattern of results from national and international assessments suggests a very wide disparity between rural and urban educational districts, private and public schools, selected public schools, and boys and girls. A recent study also suggests that schools facing difficult circumstances do much less well (De Lisle et al., 2007). The central

administrative office of the MOE certainly does not have the capability to redress the weaknesses of every individual school or district. In any case, even if it had the needed human and physical resources, it would not want to intervene in such a way, because such over-involvement might lead to further disempowerment. Greater collaboration is certainly necessary between different Divisions of the MOE. Certainly, planning for systemic reform requires joint planning by the DCD, the Student Support Services, the Division for School Supervision, and the DERE. The requirement for collaborative work is a major hurdle, but without it, compensatory education reforms are difficult to implement.

Empowering teachers to use data better requires building assessment and data literacy. Currently, the low levels of teacher assessment literacy will lead to poor classroom assessment practice. Moreover, it portends possible data misuse. An example of this problem is the tendency for teachers to mimic large-scale pencil and paper tests rather than to use performance standards to guide and develop assessments that promote classroom learning. Therefore, while the introduction of performance standards might be a significant advance in the evolution of national assessment in Trinidad and Tobago, it will not by itself lead to systemic improvement in the education system. While performance standards are necessary for evaluating quality, only a comprehensive and enacted policy on data use will ensure systemic improvement. Therefore, the MOE must move towards developing a national evaluation system with clear policies for assessments and data use. Successful implementation of such a system would require cooperation of different arms of the MOE working together to tackle achievement gaps that now exist and those that might emerge (Porter, 2005).

Declaration of Competing Interests

The author is the standard-setting consultant for the primary school national assessments of educational achievement (2004–2007).

Acknowledgements

The author wishes to acknowledge the independent work of officers of the DERE in developing and implementing the national assessments of educational achievement. The officers have always been keen to provide feedback and are intimately involved in the management of the standard-setting exercise. The author wishes to single out Mr. Peter Smith, Mr. Meryvn Sambucharan, and Mr. Harrilal Seecharan for their assistance.

References

Anderson, J. A. (2005). *Accountability in education* (Education Policy Studies). Paris: The International Institute for Educational Planning; and Brussels: International Academy of Education.

Baird, J-A. (2007) Alternative conceptions of comparability. In P. Newton, J-A Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 124–165). London: Qualifications and Curriculum Authority.

Benveniste, L. (2002). The political structuration of assessment: Negotiating state power and legitimacy. *Comparative Education Review, 46*(1), 89–118.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137–172.

Bhola, D., Impara, J., & Buckendahl, C. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues & Practice, 22*(3), 21–29.

Braithwaite, R. H. E. (1981). Plus ca change. *Trinidad and Tobago Education Forum, 2*(1), 5–12.

Brandon, P. R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education, 17*(1), 59–88.

Brandon, P. R. (2005). Using test standard-setting methods in educational program evaluation: Addressing the issue of how good is good enough. *Journal of Multidisciplinary Evaluation, 2*(3), 1–29.

Braun, H. & Kanjee, A. (2006). Using assessment to improve education in developing nations. In H. Braun, A. Kanjee, E Bettinger, & M. Kremer (Eds.), *Improving education through assessment, innovation, and evaluation* (pp. 1–46). Cambridge, MA: American Academy of Arts and Sciences.

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

Broadfoot, P., & Black, P. (2004). Redefining assessment? The first ten years of assessment in education. *Assessment in Education: Principles, Policy and Practice, 11*(1), 7–26.

Carson, J. D. (2001). Legal issues in standard setting for licensure and certification. In G. J. Cizek (Ed.), *Setting performance standards; Concepts, methods, and perspectives* (pp. 427–444). Mahwah, NJ: Lawrence Erlbaum.

Chapman, D. W., & Snyder, C. W. (2000). Can high stakes national testing improve instruction: Reexamining conventional wisdom. *International Journal of Educational Development, 20*(6), 457–474.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement, 30*(2), 93–106.

Cizek, G. J. (1995, April*). Standard setting as psychometric due process: Going a little further down an uncertain road.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Crespo, M., Soares, J. F., & de Mello e Souza, A. (2000). The Brazilian national evaluation system of basic education: Context, process and impact. *Studies in Educational Evaluation, 26*, 105–125.

De Lisle, J. (2005). *Standard setting manual for the Trinidad and Tobago Ministry of Education's standard setting exercise for the 2005 national assessments*. Unpublished manuscript, St. Augustine, Trinidad: Centre for Medical Sciences Education, Faculty of Medical Sciences, UWI.

De Lisle, J., Lewis, Y., Mc David, P., Smith, P., Keller, C., Jules, C., Lochan, S., & Seunarinesingh, K. (2007). *In the context of Trinidad and Tobago, how do we identify schools that are succeeding or failing amidst complex and challenging circumstances?* Paper presented at the Biennial Conference of the UWI Schools of Education: Reconceptualising the Agenda for Education in the Caribbean, 23–27 April, 2007 at the School of Education, Faculty of Humanities & Education, UWI, St. Augustine

Downing, S. M., & Haladyna, T. M. (Eds.) (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Earl, L. M., & Katz, S. (2006). *Leading schools in a data-rich world: Harnessing data for school improvement.* Thousand Oaks, CA: Corwin.

Eisemon, T. O. (1990). Examination policies performance to strengthen primary schooling in African countries. *International Journal of Educational Development, 10*, 69–88.

Elley, W. B. (1992). *How in the world do students read?* The IEA Study of Reading Literacy. Hamburg: International Association for the Evaluation of Educational Achievement.

Evans, H. (2001). *Inside Jamaican schools*. Mona, Jamaica: University of the West Indies Press.

Ferrer, G. (2006). *Educational assessment systems in Latin America: Current practice and future challenges*. Washington, DC: Partnership for Educational Revitalization in the Americas.

Firestone, W. A., Schorr, R. Y., & Monfils, L. (Eds.). (2004). *The ambiguity of teaching to the test*. Mahwah, NJ: Lawrence Erlbaum.

Fiske, E. (2000). *Education for all: Status and trends 2000: Assessing learning achievement*. Paris: UNESCO.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18*(3), 223–232.

Gomez, J., & Padilla, J. L. (2004). The evaluation of consequences in standards-based test score interpretations. *Measurement: Interdisciplinary Research & Perspectives*, *2*(2), 104–128.

Grant, N. (2006). Teacher resistance in the United Kingdom. *Rethinking Schools Online, 20*(4). Retrieved February 21, 2006, from http://www.rethinkingschools.org/archive/20_04/uk204.shtml

Greaney, V. (1996). Stages in national assessment. In P. Murphy, V. Greaney, M. E. Lockheed, & C. Rojas (Eds.), *National assessments: Testing the system* (pp. 111–128). Washington, DC: World Bank.

Greaney, V., & Kellaghan, T. (1996). *Monitoring the learning outcomes of education systems*. Washington, DC: World Bank.

Greaney, V., & Kellaghan, T. (2007). *Assessing national achievement levels in education*. Washington, DC: World Bank.

Green, D. R., Trimble, C. Scott, & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues & Practice*, *22*(1), 22–32.

Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco, CA: Jossey-Bass.

Gvirtz, S., & Larripa, S. (2004). National evaluation system in Argentina: Problematic present and uncertain future. *Assessment in Education*: *Principles, Policy and Practice*, *11*(3), 349–364.

Haertel, E. H., & Lorie, W. A. (2004). Validating standards-based score interpretations. *Measurement: Interdisciplinary Research and Perspectives, 2*(2), 61–103.

Haladyna, T. M. (2006). Roles and importance of validity studies in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 739–758). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., Brennan, R. L., Brown, W., Dodd, B., Forsyth, R. A., Mehrens, W. A., Nellhaus, J., Reckase, M., Rindone, D., van der Linden, W. J., & Zwick, R. (2000). A response to "Setting reasonable and useful performance standards" in the National Academy of Sciences' "Grading the nation's report card." *Educational Measurement: Issues and Practice, 19*(2), 5–14.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Washington, DC: American Council on Education.

Hansche, L. N. (1998). *Handbook for the development of performance standards*: *Meeting the requirements of Title I.* Washington, DC: United States Department of Education.

Hanushek, E. A., & Wobmann, L. (2006). Does educational tracking affect performance and inequality? Differences-in-differences evidence across countries. *Economic Journal, 116*(510), C63–C76.

Harvey, C. (1981). *Practitioner's perceptions of an innovative school system in a developing country: A qualitative analysis*. Unpublished doctoral dissertation, University of Toronto, Canada.

Hazel backs down, SEA exam here to stay. (2006, July 9). *Trinidad & Tobago Express*. Retrieved February 21, 2006, from http://www.trinidadexpress.com/index.pl/article_archive?id=160980190

Herman, J. L., & Haertel, E. H. (Eds.). (2005). *Uses and misuses of data for educational accountability and improvement.* Chicago, IL: National Society for the Study of Education.

Heubert, J. P., & Hauser, R. M. (Eds.). (1999*). High stakes: Testing for tracking, promotion, and graduation.* Washington, DC: National Academies 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.

Hiemstra, R. (2001). Uses and benefits of journal writing. In L. M. English & M. A. Gillen, (Eds.), *Promoting journal writing in adult education* (New Directions for Adult and Continuing Education, No. 90; pp. 19–26). San Francisco, CA: Jossey-Bass.

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.

Jaeger, R. M., & Mills, C. (2001). An integrated judgment procedure for setting standards on complex, large-scale assessments. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 313–318). Mahwah, NJ: Lawrence Erlbaum.

Jones, L. V. (2001). Assessing achievement versus high stakes testing: A crucial contrast. *Educational Assessment, 7*(1), 21–28.

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64,* 425–461.

Kane, M. T. (1998). Choosing between examinee-centered and test-centered standard setting methods. *Educational Assessment, 5*(3), 129–145.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kellaghan, T., & Greaney, V. (2001). The globalisation of assessment in the 20th century. *Assessment in Education, 8*(1), 87–102.

Kellaghan, T., & Greaney, V. (2004). *Monitoring performance: Assessment and examinations in Africa.* Paris: Association for the Development of Education in Africa; Washington DC: World Bank.

Lewis, S. (2005). Issues related to disaggregating data in a new accountability. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era*. Mahwah, NJ: Lawrence Erlbaum.

Linn, R. (2005). Issues in the design of accountability systems. In J. Herman & E. Haertel (Eds.), *Uses and misuses of data for educational accountability and improvement* (National Society for the Study of Education Yearbook, Vol. 104, Pt. 2: pp. 78–98). Chicago: NSSE.

Manning, H. (2004). *Senate contribution on the 2003-2004 Appropriation Bill*. Retrieved February 21, 2006, from http://www.ttparliament.org/hansard/senate/2004/hs20041025.pdf

Mayo, C. (2005) Testing resistance: Busno-cratic power, standardized tests, and care of self. *Educational Philosophy and Theory, 37*(3), 357–363.

McDonnell, L. M. (2005). Assessment and accountability from the policymakers' perspective. In J. L. Herman & E. H. Haertel (Eds.), *Uses and misuses of data in accountability testing* (Yearbook of the National Society for the Study of Education, Vol. 104, Pt. 1; pp. 35–54). Boston, MA: Blackwell.

McGinty, D. (2005). Illuminating the "black box" of standards setting: An exploratory qualitative study. *Applied Measurement in Education, 18*(3), 269–287.

Meara, K. C., Hambleton, R. K., & Sireci, S. G. (2001). Setting and validating standards on licensure and certification exams: A survey of current practices. *CLEAR Exam Review, 7*(2), 17–23.

Meyer, L. (2005). *A basic guide to standard setting.* London: AQA. Retrieved October 21, 2006, from http://www.aqa.org.uk/over/pdf/guidetostandard setting.pdf

Miyazaki, I. (1976). *China's examination hell: The civil service examinations of Imperial China.* New York: Weatherhill.

Muijs, D., Harris, A., Chapman, C., Sroll, L., & Russ, J. (2004). Improving schools in socioeconomically disadvantaged areas – A review of research evidence. *School Effectiveness and School Improvement, 15*(2), 149–175.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 International Report: IEA's Progress in International Reading Literacy Study in Primary Schools in 40 Countries.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Nitko, A. J. (1980). Distinguishing the many varieties of criterion-referenced tests. *Review of Educational Research, 50*(3), 461–485.

Oakland, T. (1995). Test use with children and youth internationally: Current status and future directions. In T. Oakland & R. Hambleton (Eds.), *International perspectives on academic assessment* (pp. 1–24). Boston, MA: Kluwer.

Olivares, J. (1996). Inclusive national testing: Chile's 'quality of education assessment system.' In A. Little & A. Wolf (Eds), *Assessment in transition: Learning, monitoring and selection in international perspective* (pp. 118–133). Oxford, Pergamon.

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

Perry, G. E., Arias, O. S., Lopez, J. H., Maloney, W. F., & Serven, L. (2006*). Poverty reduction and growth: Virtuous and vicious circle*s. Washington, DC: World Bank.

Plake, B. S. (2008). Standard setters: Stand up and take a stand! *Educational Measurement: Issues and Practice, 27*(1), 3–9.

Plake, B. S., & Impara, J. C. (2001). Ability of panelists to estimate item performance for a target group of candidates: An issue in judgmental standard setting. *Educational Assessment, 7*(2), 87–97.

Porter, A. C. (2005). Prospects for school reform and closing the achievement gap. In C .A. Dwyer (Ed.), *Measurement and research in the accountability era* (pp. 59–95). Mahway, NJ: Lawrence Erlbaum.

Ravela, P. (2005) A formative approach to national assessments: The case of Uruguay. *Prospects, 35*(1), 21–43.

Reckase, M. (2000). A survey and evaluation of recently developed procedures for setting standards on educational tests. In M. L. Bourque & S. Byrd (Eds.), *Student performance standards on the National Assessment of Educational Progress: Affirmation and improvements* (pp. 41–70). Washington, DC: National Assessment Governing Board.

Schafer, W. D. (2005). Criteria for standard setting from the sponsor's perspective. *Applied Measurement in Education, 18*(1), 61–81.

Sebatane, M. (2000). International transfers of assessment: Recent trends and strategies. *Assessment in Education: Principles, Policy and Practice, 7*(3), 401–416.

Shapiro, J., & Trevino, J. M. (2004). *Compensatory education for disadvantaged Mexican students: An impact evaluation using propensity score matching* (World Bank Policy Research Working Paper No. 3334). Washington DC: World Bank.

Skorupski, W., & Hambleton, R. K. (2005). What are panelists thinking when they participate in standard setting studies? *Applied Measurement in Education, 18*(3), 233–256.

Tomlinson, M. (2002). *Inquiry into A Level standards – final report.* London: HMSO.

Trinidad and Tobago Chamber of Industry and Commerce. (2006, July 13). Let's harness our nation's intelligentsia. *Business Guardian,* p. 29.

Trinidad and Tobago. Ministry of Education. (2005). *SEA report: Secondary Entrance Assessment progress report*. Port of Spain, Trinidad: Author.

Trinidad and Tobago. Ministry of Education. Division of Educational Research and Evaluation. (2005). *Proposal for standard setting: The 2005 national test.* Port of Spain, Trinidad: Author.

Trinidad and Tobago. Ministry of Education. Division of Educational Research and Evaluation. (2006). *National test report, 2005.* Port of Spain, Trinidad: Author

Wall, D. (2005). *The impact of high-stakes testing on classroom teaching: A case study using insights from testing and innovation theory* (Studies in Language Testing, volume 22). Cambridge, UK: Cambridge University Press.

Wilbrink, Ben (1997). Assessment in historical perspective. *Studies in Educational Evaluation, 23*(1), 31–48.

Winkler, D. (2000). Educating the poor in Latin America and the Caribbean: Examples of compensatory education. In F. Reimers (Ed.), *Unequal schools, unequal chances: The challenges to equal opportunity in the Americas* (pp. 113–132). Cambridge, MA: David Rockefeller Center for Latin American Studies, Harvard University.

Wolff, L. (2004). Educational assessments in Latin America: The state of the art. *Applied Psychology: An International Review, 53*(2), 192–214.

World Bank. (1993). *Caribbean region: Access, quality, and efficiency in education.* Washington, DC: Author.

World Bank. (1995). *Trinidad and Tobago: Poverty and unemployment in an oil based economy* (Report No. 14382-TR). Washington, DC: Author.

Zieky, M. (1997). Is the Angoff method really fundamentally flawed? *CLEAR Exam Review, 8*(2), 30–33.