Journal of International Education and Leadership
http://www.jielusa.org/
Volume 2 Issue 3 Fall 2012
ISSN: 2161-7252

Assessment of GenNEXT Learning Outcomes at the University of Brunei Darussalam: A Qualitative Review of Selected Opportunities, Benefits and Challenges in Human Resource Development

Lawrence Mundia
University of Brunei, Darussalam

The commentary and overview explored how curriculum and assessment reforms are being used by a small university and small country to improve the quality of education and gain international recognition. Although the reforms are potentially beneficial to the students, university, and country, there are dilemmatic factors that may either enhance or harm the quality of the assessments. Eight problems discussed in this review report are: standards in educational testing and rights of test takers; norm-and-criterion referenced tests; relationship between formative and summative evaluations; moderation of marks for borderline students; use of optional questions in examinations; types and quality of the assessments; use of raw and standard scores in selection; and role of external assessors. Strategies to use these effectively are suggested. Lessons drawn from these initiatives might be of concern and interest to other small universities and countries. Further reviews of educational quality addressing other setbacks are encouraged and invited.

Keywords: Assessments; quality; leadership; innovation; entrepreneurship; environmental awareness.

Brunei education has undergone many changes since the country attained independence in 1985. The transformations have occurred at all levels of the education system. Often, the innovations were also accompanied by adaptations in the academic assessment procedures. Some of the major reforms are briefly described below.

Reform of the Brunei school system

In the past fifteen years (1997-2011), Brunei Darussalam has implemented three major educational policies: (1) inclusive education, in 1998; (2) the National Education System for the 21st Century known in Bahasa Melayu language as Sistem Pendidikan Negara Abad 21 or SPN21, 2008; and (3) education for the gifted / talented students, 2009. These changes were partly designed to diversify and broaden accessibility of education as well as to improve the quality of education. Following the implementation of the policy of inclusive education (Ministry of Education, 1997; 1998), examinations were adjusted to accommodate the needs of students with special needs. Brunei students with mild, moderate and high support needs now attend regular schools and write the same examination papers as their non-disabled peers. Some of the accommodations that need to be made for such students are discussed in detail by Murray (1996). Another major change undertaken was the removal or abolition of the "N" Level examinations. Under the ongoing SPN21 curriculum changes (Ministry of Education, 2007), the government of Brunei has rightly called for the reform of the school assessments and introduced new evaluations such as the Student Progress Assessment, School-Based Assessment, and Student Progress Examination. Continuous assessment (also known as formative evaluation) now contributes 30% to the final grade. Summative evaluations were retained at primary (Year 6) and junior high school (Form 3) levels. Most teachers in Brunei would be familiar with School-Based Assessment (formative evaluation) but not so with the Student Progress Assessment and the Student Progress

Examination. These various forms of continuous assessment are also known as "Check-Point Assessments" in the Brunei education system under the current curriculum reforms. To minimize or clear confusion, some schools are organizing seminars and workshops to brief teachers on the possible implications of SPN21 curriculum on school assessments and the various types of assessments that will be used (see Clark, 2009). Specifically, the government wants to change these assessments from being norm-referenced to criterion-referencing to emphasize mastery of knowledge and competency skills. But this might not be enough. To achieve the overall comprehensive objectives of the SPN21 curriculum, school summative examinations and teacher education need to be innovated in several other ways.

Reform of teacher education programs

To effectively support the ongoing SPN21 curriculum reforms and implementation of other educational policies (inclusive and gifted education), the pre-service and in-service teacher education courses have, since 2009, been modified to make them responsive to innovations and adaptations in the school system. The Sultan Hassanal Bolkiah Institute of Education (SHBIE) is one of the faculties of the University of Brunei Darussalam (UBD) charged with the responsibility of training teachers for government schools. Prior to August 2009 SHBIE prepared teachers for various undergraduate teaching qualifications (certificate in education, diploma in education, postgraduate certificate in education, and bachelor degrees in education) as well as postgraduate level credentials (master of education and doctor of philosophy in education). In the middle of 2008 the government of Brunei Darussalam, through the Ministry of Education, introduced a new policy in teacher education that required SHBIE to train teachers at only the postgraduate levels such as master of teaching (MTeach), master of education (MEd), and doctorate of philosophy in education (PhD). The MTeach degree has four specialization strands (early childhood education; primary education; secondary education; and vocational and technical education). Implementation of the new policy started effective from August 2009. From then onwards students who are interested in becoming teachers have to do an undergraduate degree in other faculties and acquire in-depth content in a subject(s) teachable in schools before they can take the MTeach degree to qualify them to teach in schools. Details of the new MTeach teacher education programs are available on the University of Brunei Darussalam website address (http://www.ubd.edu.bn). The main reason given by the government for this change is that the country wanted to raise the qualifications of teachers and thereby improve the overall quality of education in the nation. Meanwhile the status of SHBIE at UBD was raised to a graduate faculty.

Reform of university education programs

In addition to implementing teacher education reforms in the Faculty of Education and to align itself more with the country's objectives regarding the provision of quality education under SPN21 reforms, UBD has, since August 2009, embarked on an ambitious goal of providing its students a world-class education that will equip them with 21st century skills required in a rapidly changing global workplace. The university's aspirations are described in detail in a brochure titled "Universiti Brunei Darussalam - Our National University". This document is available in both print and soft formats. The online version can be accessed from the UBD official website address (http://www.ubd.edu.bn) under the title "UBD Brochure". Throughout the present study, the blueprint document will, for simplicity, be referred to as the "UBD Brochure (2012)". Besides arming its graduates with 21st century skills, the other main objective of UBD (though not specifically stated in the brochure document), is to be among the 50 top and best universities in Asia by 2015. Altogether, imparting 21st century skills and being ranked one of the top 50 universities in the Asia region of the world, constitute the core of UBD's current vision, mission, and values. To achieve these aims, the university actively supports the implementation of a revolutionary next generation (abbreviated or coded as GenNEXT) curriculum that is interactive,

interdisciplinary, research-based, technologically-mediated, learner focused, experiential-and-outcome oriented, and addresses individual learning styles. In addition, the university has identified four broad domains that it believes en-campus the desirable 21st century skills: leadership; innovation; entrepreneurship; and environmental awareness. These areas need to be incorporated, not only in the university curriculum, but also among the student assessment and evaluation strategies to foster or facilitate the teaching, studies and research in these fields. Although the curriculum, teaching and assessment are all important interrelated activities, the present review is neither an evaluation of the GenNEXT curriculum and teaching issues nor an evaluation of the assessment of GenNEXT skills. These developments are too recent at UBD to be accorded a realistic and meaningful evaluation at the present time. In view of this, these matters are outside the objectives and beyond the scope of the present review. Instead, the purpose of the present review is to make few personal comments or personal viewpoints on three issues: (a) the opportunities that the changes in academic assessments may bring in improving the quality of education provided by UBD; (b) the benefits of improved education to the university, students, academic staff and the country; and (c) to discuss a few selected challenges and setbacks that might be encountered in the process of implementing assessment innovations. Each of these issues is separately and briefly discussed below. Furthermore, in the context of the present review, the term "assessment" operationally refers to tests, exams and other alternative procedures for measuring and evaluating learning, development, and growth achieved by students as a result of attending the university.

Opportunities for improving the quality of education

As indicated in Fig 1 below, there is clearly a cyclical relationship between the curriculum, teaching and assessment.

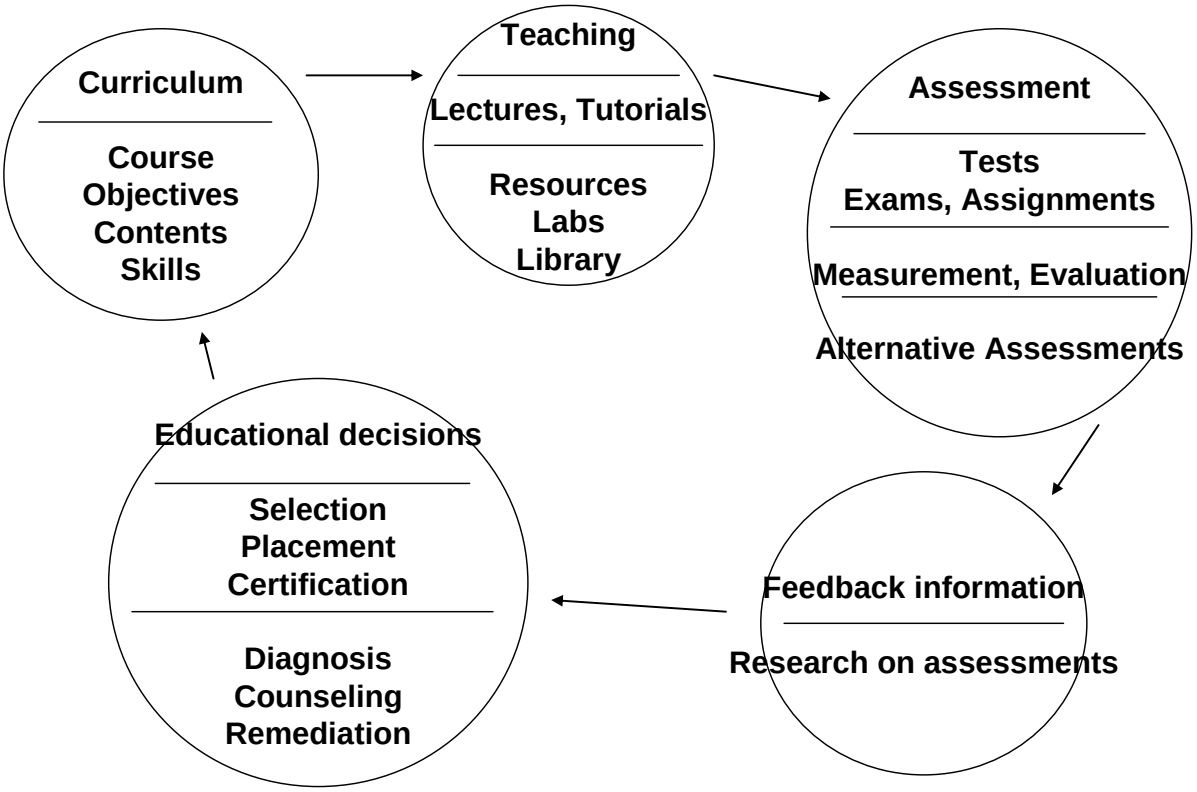

Figure 1: Cycle of curriculum, teaching and testing objectives

This suggests that mere inclusion of content domains such as leadership, innovation, entrepreneurship, and environmental awareness in the curriculum for different courses (units or modules) may, in itself, not necessarily result in students' mastery of skills or competencies embedded in these areas. Unless the salient and desired skills from these fields are incorporated among the academic assessment instruments, it appears that teaching and learning might not emphasize them. Educational assessments are well known for being powerful sources of extrinsic motivation in teaching and learning. Contents and skills in the curriculum that are not frequently examined are often ignored or overlooked by teachers and students. The implication drawn from all this is that UBD academic staff now have an excellent opportunity to be creative, innovative, adaptive and resourceful in designing high quality assessments (both cognitive and non-cognitive) that tap the 21st century skills such as critical thinking, application, analysis, synthesis, creativity, problem solving, evaluation, determining plausible implications, computer literacy, and making intelligent predictions in real-life environments. Critics and skeptics of these UBD academic reforms might ask the question: will it be possible or feasible to assess leadership, innovativeness, entrepreneurial, and environmental awareness skills in all courses taught at UBD? To a large extent, the answer to this question is "yes" (given the teaching and learning resources available to instructors and students at UBD as well as the level of moral support accorded to these stakeholders). And as pointed out above, the inclusion and implementation of these 21st century

skills in the curriculum and assessments, may greatly improve the quantity and quality of education offered at UBD. A brief mention of the specific target skills and available supportive resources and facilities (financial, material and human) is worth making here. Listed and briefly discussed below are the key skills emphasized by UBD in the current curriculum and assessment reforms.

Leadership and innovation (UBD Brochure, 2012 pp. 8; 9; 17). Living in an era of breathtaking and accelerating change, students' learning will be enhanced through the development of life and career skills that encompass:

- flexibility and adaptability
- initiative and self-direction
- social and cross-cultural skills
- productivity and accountability
- leadership and responsibility

The teaching and assessment of the above contents and skills will be facilitated in a variety of ways. For example, students will benefit from the low student-to-lecturer ratio at UBD, high caliber of the teaching staff (both local and foreign), a well-resourced library, and excellent laboratories for various disciplines such as mathematics, science, language, computing/ICT, and medicine. In addition, the Institute for Leadership, Innovation, and Advancement, ILIA (founded in 2009) offers short and CPD hands-on-experience workshop training programs in leadership, negotiation, decision-making strategies, strategic change, and management for academic staff, public service employees, and workers in the private sector. Other short courses and CPD workshops that support academic staff are conducted under the Teaching and Learning Centre (TLC) within UBD.

Entrepreneurship (UBD Brochure, 2012 pp. 8; 9; 17). In an evolving global enterprise era, UBD is poised to produce a labor force that is creative, resourceful, adaptable, and innovative. GenNEXT's curriculum and assessment emphases on entrepreneurship are deliberately intended to enable students to develop the following attributes needed in responding quickly:

- seek opportunities
- act quickly
- negotiate
- build networks

The UBD Innovation and Enterprise Office (IEO) provides information and advice on the commercialization of intellectual property rights. The university recognizes that researchers, organizations and enterprises work best in an environment that enables them to protect their competitive advantages and be rewarded for their contributions to new products, services and developments in technology. The IEO aims to facilitate academics, students, researchers, entrepreneurs and organizations in obtaining intellectual property rights, including patents, trademarks, copyrights and trade secrets, for their creativity and innovations.

Environmental Awareness (UBD Brochure, 2012 pp. 9; 17).UBD is committed to ensuring that every member of the university is environmentally aware and responsive to serious global environmental issues, including climate change, global warming, ozone layer depletion and over-exploitation of natural resources. UBD supports and conducts several internationally relevant and urgent research projects which seek to address some of these issues (UBD Brochure, 2012). Currently, two research groups (Energy and Biodiversity), one research cluster (Environment and Sustainable Development) and one research centre (Kuala Belalong Field Studies Centre) are

actively involved in probing the environment and bringing its awareness to members of the public. Under the GenNEXT programs, experiential learning is seen as the main way for exposing students and enabling them to explore the environment and the real-world of work. Students will be moved out of their classroom environment to undergo a discovery year to gain community-based or international experience outside the UBD campus. Sensitization of the students about the environment might be achieved by asking students to choose one of the following options or in combination:

- Enroll in a study abroad program at one of UBD's partner universities overseas.
- Take an internship, practicum, or work attachment with an appropriate company, agency, or NGO.
- Plan, implement, and evaluate a community-based outreach program.
- Create a start-up business or an innovation enterprise.

Design of GenNEXT academic assessments

UBD offers a wide range of courses and programs, all of which are now required to integrate the 21st century skills discussed above. While it may be easy to include these skills in the assessment instruments, it would be unrealistic to expect or assume that the development and construction of GenNEXT tests/examinations will follow the same format (or be standardized) across the faculties and departments. Below are a few examples of how assessment items might be prepared. Items may differ in difficulty and sophistication depending on the subject, type of assessment (e.g. objective test, essay test, problem/performance-based assessment, experiment, or research project), level of assessment or award (e.g. certificate, diploma, bachelors, masters, and doctoral), and time available for students to think through and write the responses in the required format. The items would need to have a high level of content validity demonstrated through content analysis and constant comparison by expert judges. Note that the examples provided here are only for illustration purposes. The programs, departments and faculties mentioned below are fictitious or imaginary and do not exist at UBD.

Exhibit # 1: Leadership styles - Educational Psychology; Teacher Education Program; Faculty of Education.

"Suppose three different math classes of mixed-ability "O" Level students with different special needs are being taught by three different types of teachers: autocratic; authoritative; and laissez-faire. During a math lesson, the teacher works out on the whiteboard a solution to an algebra problem but makes some errors. One of the gifted/talented students corrects the teacher in the presence of all other students". Imagine and briefly describe how each of these teachers might react to the following hypothetical scenarios in a classroom situation.

(a) What would each of the teachers say to the more able student? Give reasons for your answer.
(b) What would each teacher say to the other students? Give reasons for your answer.
(c) What would each teacher do to the math problem? Give reasons for your answer.
(d) In which ways would the responses of the autocratic and laissez-faire teachers differ? Why?
(e) Giving reasons for your answer, briefly explain the likely effect of each teachers' leadership style on the students'
    - initiative and creativity

- motivation
- self-esteem
- self-efficacy in mathematics

Comment: this item is too long. It might be suitable for a term paper.

Exhibit # 2: Innovative practices. Fine Arts; Design and Technology Program; Faculty of Social Sciences.

Think of an object people like using most in the home environment e.g. a plastic cup or metal spoon.

(a) List all the uses of the cup or spoon.
(b) In which ways can the cup/spoon be transformed to make another useful object?
(c) What are the uses of each derived object?
(d) How else may materials in the cup/spoon be recycled to make other goods?

Comment: this item is neither too long nor too short. It might be suitable for a test or an examination.

Exhibit # 3: Entrepreneurial skills. Sales logistics; Marketing Research Program; Faculty of Business Studies.

Design an advertisement for selling a company product that will target a given audience through a given publication outlet. Select the product, audience, and publication outlet. Explain the process of design and the choices you made in creating the advert. Finally, develop rubric dimensions for marking this project.

Comment: this item is too long and suitable for a project format of assessment.

Exhibit # 4: Environmental awareness. Environmental hazards; Environment Protection Program; Faculty of Science.

Situation/Scenario

Imagine you are living in a highly toxic and health-risky environment. The air is heavily polluted by industrial carbon fumes and vehicle emissions. River water and sea water on the beaches are infested and choked with garbage, sewage waste, and industrial residues. Underground water is contaminated and threatened by acidic spills from chemical plants.

Assessment task
Based on the above information, design either a community-wide (or school-wide, or worksite-wide) message campaign with the following purposes:

(a) To sensitize the target group about the health risks associated with living in such an environment.
(b) To brainstorm ways and means by which environmental pollution may be reduced and eventually prevented.
(c) To plan, implement, and evaluate one small-scale project intended to reduce environmental degradation.

To ensure that your campaign is effective, clearly suggest and explain what should be done to the following factors:

- Source of the message
- Type of message disseminated
- Medium used to send the message
- Target group/audience characteristics
- Situation/context under which the message is distributed

Comment: this item is too long, big and ambitious. It may be suitable for an action research project or a quasi-experiment, involving both the researcher(s) and members of the target group. Furthermore, the item might be appropriate for a practicum, an internship or work attachment format and setting.

Benefits to students, academic staff, UBD and Brunei

The introduction and implementation of GenNEXT curriculum and assessments at UBD is expected to benefit a wide range of stakeholders. The anticipated benefits to students, academic staff, the country, and UBD itself are briefly discussed below.

Potential benefits to students. The skills embedded in GenNEXT curriculum and assessments are thought to be extremely beneficial to UBD students in that they make university education highly relevant by tailoring it to the country's skill needs. As already pointed out in the discussion above, both the Brunei government (Ministry of Education) and UBD are at present concurrently and concordantly talking of 21st century education and skills via SPN21 and GenNEXT curriculum reforms, respectively. With the implementation of these reforms (which are based on extensive consultations with both local and foreign experts), university students may no longer take irrelevant courses or degree programs for which there are no jobs on the labor market. In addition, Brunei might reduce its dependence on expatriate personnel in key jobs, particularly the technical ones. If these government and UBD reforms are successful, then gone are the days when UBD graduates used to spend months or years looking for gainful employment and welcome to the economic phase of more employment opportunities for Brunei nationals.

Anticipated benefits for academic staff, UBD and Brunei. Under the ongoing GenNEXT programs, teaching and assessment are supposed to be research-based to a large extent. This therefore implies that academic staff must be active, proficient and vibrant in conducting, publishing and using research. In UBD's vision (to be 1st class international university) and mission (to be among the top 50 universities in Asia by 2015), the volume of research (both quantity and quality) per academic staff, is considered to be an important key performance indicator (KPI). Research output is one of the many and diverse criteria used by organizations/institutions such as the Shanghai Jiao Tong University (Academic Ranking of World Universities, see http://www.ed.sjtu.edu.cn/ranking.htm) and the Times Higher Education Supplement (World University Rankings, see http://www.thes.co.uk/worldrankings/) that rank universities along some quality dimensions. Indicators of research quality include (but are not exclusively restricted to): the number of reviewed journal articles, books, monographs and other publications produced by each academic staff per year; number of articles published in high impact factor journals such as Tier 1 and Tier 2 journals; number of citations per faculty member; number of obtained research grants used, and the number of patents attained. UBD academics have repeatedly been told at different forums (meetings, seminars and workshops) that research will be weighted high or heavily in decisions concerning staff appointments, annual appraisal, contract renewal, and promotion. The present KPI per academic staff in all UBD faculties is each instructor to publish at least two articles annually in Tier 1 and Tier 2 reviewed journals. These measures may stimulate and motivate lecturers to produce meaningful/realistic research that is useful in solving local and international problems. Such research might also be incorporated in teaching

thereby improving the overall quality of education. With emphasis on internationalization of the university, highly experienced and well published professors from other countries would be encouraged to take up academic positions at UBD thereby strengthening further the quality of education and research culture and enhancing the university's rank in the world. These improvements and achievements in education might also help attract foreign students to UBD. Eventually, Brunei (currently well known as the abode of peace in Southeast Asia) may also become the hub of excellence in teaching and learning. By extrapolation, the investments in the knowledge industry would not only be a good way of diversifying Brunei's economy from oil and gas but also benefiting the country financially through foreign student fees.

Challenges in implementing GenNEXT assessments at UBD

Just like the opportunities and benefits discussed above are many, the practical challenges, dilemmas, and controversies arising from implementing the GenNEXT assessments are also equally many. This section of the review will look at eight selected implications. This list of issues is not exhaustive but rather illustrative of the wide range of concerns that will need to be addressed. These are: (1) standards in educational testing and rights of test takers; (2) norm-referenced and criterion-referenced tests; (3) relationship between formative and summative evaluations; (4) moderation of marks for borderline students; (5) caution on the use of optional questions in examinations; (6) types and quality of the assessments; (7) use of raw and standard scores in selection; and (8) role and use of external assessors. Each of these is, in turn, briefly described below.

Standards in educational testing and the rights of test takers

One of the important documents that educational testers in Brunei may need to be aware of is the Standards for Educational and Psychological Testing developed collaboratively by the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) to "promote the sound and ethical use of tests and to provide a basis for evaluating the quality of testing practices" (AERA, APA, & NCME, 1985; 1999). The Standards provides criteria for the "evaluation of tests, testing practices, and the effects of test use" (AERA, APA, & NCME, 1985; 1999). It is intended to be used by individuals and organizations that develop tests, administer tests, and use test results to make inferences about individuals or groups. The Standards covers a wide range of broad issues in testing including: test construction, evaluation, and documentation; fairness in testing; rights and responsibilities of test takers; testing individuals with disabilities; testing applications; and testing in program evaluation and public policy.

The Code of Fair Testing Practices in Education (hereafter known as the Code) is a supplement to the Standards discussed above and was prepared by the Joint Committee on Testing Practices, JCTP, a cooperative effort among several professional organizations, in 1988. It has since then been revised twice (in 2000 and 2004). The Code is a guide for professionals in fulfilling their obligation to provide and use tests that are fair to all test takers regardless of age, gender, disability, race, ethnicity, national origin, religion, sexual orientation, linguistic background, or other personal characteristics. Fairness is a primary consideration in all aspects of testing. The Code applies broadly to testing in education (as an aid admissions, educational assessment, educational diagnosis, and student placement) regardless of the mode of presentation. It is relevant to conventional paper-and-pencil tests, computer-based tests, and performance tests. It is not designed to cover employment testing, licensure or certification testing, or other types of testing outside the field of education. Although the Code is not intended to cover tests prepared by teachers for use in their own classrooms, teachers are encouraged to use the guidelines to help improve their testing practices. The aim of the Joint Committee is to act, in the public interest, to advance the quality of testing practices. Members of the Joint Committee include the American

Counseling Association (ACA), the American Educational Research Association (AERA), the American Psychological Association (APA), the American Speech-Language-Hearing Association (ASHA), the National Association of School Psychologists (NASP), the National Association of Test Directors (NATD), and the National Council on Measurement in Education (NCME).

Besides the Standards and the Code discussed above, educational testers in Brunei might also wish to know more about available training opportunities in educational testing and the qualifications that can be obtained. Literature is abundant on training and qualifications in educational testing (e.g. Eyde et al., 1988; 1993; Moreland et al., 1995; Hambleton, 1994; Bartram, 1995; 1996; and Joint Committee on Testing Practices, 1988; 2000; 2004). The British Psychological Society (BPS) offers professional training in educational and psychological testing at two levels (A and B): Certificate of Competence in Educational Testing, CCET, Level A; and Certificate of Competence in Psychological Testing, CCPT, Level B. Boyle and Fisher (2008) explain in detail the BPS standards in educational and psychological testing as well as Levels A and B training courses.

Assessment contract, ethical obligations and legal implications

With due emphasis on coursework and the use of a variety of evaluation strategies under GenNEXT curriculum and assessments at UBD, students may wish to have more rights and take more responsibility for their learning and assessment. According to Ory and Ryan (1993) students have a right to know how they might be assessed on a course. Ory and Ryan suggest that the instructor and students ought to mutually agree early on the course assessment strategies. Changes to agreed assessment procedures should be communicated to students on time and mutually agreed upon by all parties concerned. Preferably, the whole task of generating, processing and keeping marks should be transparent. One of the most frequently asked questions by UBD students on continuous assessment is: can I know my coursework marks and grade? The question demonstrates and illustrates the importance of coursework assessment which currently ranges from 40-100% for some modules at UBD. Instructors should not take this question lightly. It has lots of practical, ethical, and legal implications. This question might be addressed using a form like the one suggested in Table 1 below. Each student can complete, sign and date this form. The signature appended on this form would be an important legal attestation that the student has sighted the component marks, knows, and agrees about how his/her marks were generated and linearly combined without manipulation to get the composite or overall score and grade on coursework. Sighting and verifying marks and signing to confirm correctness would reduce or eliminate suspicions, complaints, appeals, and even court litigations concerning accusations of biases and favoritism in the assessment system from critics and skeptics (who might include students and their parents). In 2003 the Faculty of Education at the University of Swaziland (UNISWA) in Swaziland (a small kingdom in southern Africa about the size of the Sultanate of Brunei Darussalam) was using a variation of this form in an effort and attempt to be transparent with sensitive examination assessments. Each student was required to go to the course lecturer's office individually towards the end of the semester in order to sight and verify his/her marks and sign a document to confirm this process. The verification system actually stimulated and motivated students to work harder to improve performance. In the example below, coursework contributes 40% of the marks to the final composite score and grade. The range of grades and total percentage scores for coursework assessments at UBD is as follows: F (0-39%); E (40-49%); D (50-59%); C (60-69%); B (70-79%); and A (80-100%). The final course examination contributes 60% to the overall course marks and grades. Final examination marks are coded and categorized using the same grading scale for coursework marks indicated above. Under the present GenNEXT circumstances at UBD, recordkeeping and custodianship of students' coursework marks may assume increased importance and are likely to become contentious. Students might, for example, demand the right to know how different components of continuous assessment marks or

coursework marks would be weighted, scaled and linearly combined to get composite scores and grades especially in those modules/units that are almost (70-100%) based on course work. In general, students' marks are highly confidential and sensitive personal data that cannot be given to unauthorized persons or agencies. The marks cannot also be manipulated without the student's knowledge as it is immoral, unethical, illegal, and unprofessional to do so.

Table 1
Linear combination of coursework marks

A. Registration Number: _____
   Name: _____
B. Weighting/Scaling of coursework component marks

Test	67% x .15 = 10.05
Essay	81% x .18 = 14.58
Presentation	73% x .07 = 5.11
Total coursework marks	= 29.74

C. Coursework percentage score = 29.74/40 x 100 = 74%
D. Coursework grade: B

E. Signature: _____
F. Date: _____

Norm-referenced and criterion-referenced tests

These are two ways of interpreting test scores (Mundia, 1999). Norm-referenced test (NRT) scores can be interpreted like criterion-referenced test (CRT) scores and vice versa. Thus although NRT and CRT serve different functions, it is quite possible that a single test can be used for both selection (NRT) and certification (CRT) purposes. A single test which combines NRT and CRT functions is called a composite test. The assessments are therefore complementary rather than antagonistic (William, 1993). Below, the two types of assessment are distinguished further to help determine the implications of using them at UBD as part of the GenNEXT assessments.

Norm-referenced tests measure and compare a person's competency with that of others in the normalization or standardization sample. The norms (average characteristics) are often reported in many different ways (e.g. mean, mode, median, and standard deviation). To facilitate comparison in performance, raw scores are usually transformed into standard scores. For example, standard scores enable a student's performance to be compared with the average performance of an appropriate age-group or cohort in the reference population (William, 1991). Specific comparisons could be made using Z-scores and the table of areas under the normal curve to determine percentile ranks.

Other descriptive statistics such as deviation quotients and grade-point average (GPA) scores could also be used to rank students. Alternatively, the norms could be given as a table of the average scores obtained by each age-group or cohort in the standardization population (William, 1991). Such a table of age-norms would enable the student's score to be expressed as an age-equivalent (e.g. mental age and reading age). Developmental norms (e.g. a child's milestone achievements in sitting, walking, and talking) are examples of age-norms. Before using a standardized NRT assessment instrument, it is better to ensure that the test is evaluated against three important considerations: (1) recent and appropriate norms; (2) size and representativeness of the normalization sample; and (3) relevancy of the test to the new cohort and the curriculum. There are many reasons why NRTs are useful assessments. For instance, the teacher's knowledge of a

student's limitations and ways for helping the learner are usually enhanced by drawing on career or life experiences with similar students in the previous cohorts as anchors or referents. In this sense, the norms of a well standardized NRT assessment would normally be based on the performance of a much larger group of students thereby extending the teacher's experience and basis against which exceptional performance (too high or too low) would be assessed. These constant comparisons in educational testing form the basis for standardized admission or selection tests such as SAT, TOEFL, GRE, and GMAT. Such tests are not only reliable but have also good predictive validity. A student who passes them is, if admitted, likely to complete university studies successfully. Thus NRT assessments are not necessarily bad when compared to other assessments such as CRTs. Outside education, the whole of psychological assessment, a specialization in which each psychometric instrument (questionnaire, scale, or inventory) has its own norms, is based on norm-referencing or comparison with a reference group. The tests are used extensively in psychotherapy, counseling and psychiatry. Similarly, most of the clinical tests (e.g. temperature, weight, blood, urine, and stool) in medicine are also norm-referenced. The results are often interpreted as either falling within the normal range or as abnormal/critical (if occurring far above or below the population average). Under the ongoing SPN21 and GenNEXT assessment reforms, Brunei (Ministry of Education) and UBD may wish to consider retaining and using some of the functions of norm-referenced tests for selection purposes (e.g. using ranks to identify candidates for admission, degree classification, scholarship award, prize giving, and making nominations for other roles). The current view in the Ministry of Education appears to be that NRTs are bad and that they should be replaced by CRTs.

On the other hand, criterion-referenced tests (CRTs) assess how well a student has mastered a given number of skills or objectives. Selected total scores (raw or derived) are designated as cut-off points (standards or criteria) by which to judge a student's competency level. Occasionally, only one cut-off point (criterion or standard) is used. For example, when making a pass/fail decision, students might be told that those who got 75% have passed the test while those who scored below this point have failed. Sometimes, multiple cut-off points (criteria or standards) are established and used in determining students' mastery/competency. This always occurs when assigning grades to students using an acceptable grading scale. A good example for the application of multiple criteria/standards from the literature is provided by Hambleton's (1982) study. The distinct advantages and popularity of CRT assessments are based on three major factors. First, the assessments are designed to measure mastery or competence rather than merely ranking and comparing the students on some quality dimension(s). Whether the performance of one student is better or worse than that of other learners is of little relevance or importance in this context. Second, the student competes with the established standard(s) or criteria rather than with other students. This type of competition is believed to be less damaging in the face of repeated failure and ultimately fosters the development of intrinsic motivation and internal locus of control. Third, the assessments are tailored and tied to the objectives of the immediate classroom teaching program with evaluation instruments set by instructors to suit their own curriculum. These are salient characteristics or features of CRTs that the Ministry of Education (SPN21 curriculum) and UBD (GenNEXT programs) might wish to explore and exploit as they assess their students. Under NRT strategies, assessments may be external (e.g. "O" Level, "AS" Level, and "A" Level in Brunei government and international schools) rather than internal.

Relationship between formative and summative evaluation

In the context of the present review, formative evaluation includes coursework assessments (also variously known generically as continuous assessment, internal assessment, or school-based assessment). These assessments are conducted throughout the term, semester or academic year as teaching progresses. Somerset (1987) identified five main advantages of continuous assessment as: (a) no single test can measure all of a student's knowledge and skills; (b) repeated testing provides

a better picture of the student's ability; (c) multiple testing guarantees experience of success, a motivating factor; (d) multiply testing permits assessment of affective and psychomotor skills; and (e) students are less tense, anxious, and stressful when doing coursework tasks. On the other hand, summative assessments are the evaluations administered to students at the end of a term, semester, or academic year to summarize students' overall academic achievements. The relationship between these two forms of assessment needs to be investigated for a number of reasons or implications, the main one being contamination and interaction between them. This will need to be studied carefully (preferably by inter-correlations) to ensure that there is no redundancy or duplication in items between formative and summative assessments. When items in the two types of assessments do not repeat each other, the inter-correlations will be negative or nearly zero. The correlations may be positive but low and non-significant. Such correlations would indicate that the curricula contents and academic skills emphasized in the two assessments were different and did not replicate each other. Technically and ideally, continuous assessment is supposed to be used in assessing practical contents and skills that are not testable in final examination formats. The probability of repeating the items could be high if, for example, both continuous assessments and the final examinations draw their items from the same item bank subject by subject and year by year (Mundia, 1996). In this instance, the inter-correlations would be positive, high, and significant but the value of such educational assessments would be zero or redundant despite the students' total scores being spuriously high. Under the ongoing SPN21 and GenNEXT assessments, coursework and final examinations will both contribute to the final composite score and grade. In view of this, Brunei and UBD will need to carefully watch and study the relationship between the two forms of assessments (formative and summative) subject by subject and year by year for signs of redundancy. Both agencies (Ministry of education and UBD) should avoid encouraging the proliferation or mushrooming of assessment items/strategies that are repetitious or duplicate each other and are in the end redundant and reductionist (not adding or assessing anything new).

Moderation of marks for borderline students

According to the Ministry of Education (2007) School-Based Assessment marks will be moderated by using weights to ensure that the continuous assessment scores awarded by various schools are comparable. In addition to this, the Ministry of Education should also consider finding a standardized procedure of moderating the marks for borderline students rather than using the arbitrary rule of thumb (looking for additional marks here and there) which may differ from one testing/selection context to another. This advice also applies to UBD under the ongoing GenNEXT programs. A standardized procedure would be useful when making uniform or fair selection decisions and would greatly benefit students with special needs in inclusive schools and the less able students at UBD whose marks often fall in the borderline region. The problem here is that a student's true score may never be known due to the effect of measurement errors. Under classical or conventional test theory, an obtained test score consists of two parts: a true score component and an error score component. This is expressed and notated algebraically as: $O = T + E$. The error component (E) represents chance changes in obtained test scores from one testing occasion to another. Theoretically (in mathematics or statistics), every score or value has an interval with an upper limit and a lower limit. To know (or estimate) where the true score might lie within the interval requires employing a statistic called the standard error of measurement (SEm) for individual scores (if the reliability of the test is known) to obtain confidence limits for a given score at a given probability level. The SEm (not to be confused with the standard error of the mean which is also abbreviated as SEm) may be regarded to be an average of squared deviations of the difference between the obtained score and the true score if the test were taken an infinite number of times (assuming no change in the students' knowledge as a result of further learning). In addition, the SEm (which gives allowance for uncertainty) may also be viewed somewhat like the

reliability of the score or test. The smaller the SEm value, the more reliable a score or test is and vice versa. Some borderline cases whose marks fall within the interval [($Z_{1.96}$) (SEm)] below the cut-off point are usually told you have failed (when they have passed) and are known as false positives (Type II error). It is like a medical doctor telling a person that "you are sick because you tested positive on a flu test, when in fact the person is healthy". This error is due to the fact that the person's actual or true score might be higher and closer to the upper limit. Alternatively and though passed/selected automatically, some cases falling within [($Z_{1.96}$) (SEm)] points above the cut-off point are called false negatives (Type I error). Again, this error is due to the fact that the person's actual or true score may be much smaller and lie close to the lower limit. Returning to a medical analogy, this is similar to a doctor telling the person that "you are well because you tested negative on a flu test, when the person is in reality sick". Thus the "true" score of any obtained point (X) lies between the interval X - [($Z_{1.96}$) (SEm)] and X + [($Z_{1.96}$) (SEm)]. Such an interval is also expressed as: $X \pm [(Z_{1.96})\ (SEm)]$. This range is called the 95% confidence interval because we use a Z-value of 1.96 (or 2) in constructing it. Similarly, a 99% confidence interval may also be determined using a Z-value of 2.58 (or 3). In both of these two intervals, the cut-off point serves as the midpoint. The upper and lower limits of these intervals are called confidence limits. When moderating marks for borderline students who either fail a test or miss a grade marginally or narrowly, it would be advisable to increase a student's marks only up to the upper limit score of the confidence interval. The SEm for individual scores may be calculated using the formula: $SEm = S\sqrt{1\text{-}Rel}$ (where S = standard deviation of test scores; $\sqrt{}$ = square root; and Rel = reliability of the test). If the reliability of the test is unknown, it may be estimated using either the Kuder-Richardson 20 (KR20) formula or the KR21 method (Kuder & Richardson, 1937). Each of these KR procedures gives a lower-bound estimate value of the reliability that can be inserted in the SEm equation. To compute the KR20 or KR21 reliability coefficient, the tester needs to know only three quantities: the number of items on the test; the mean of the test scores; and the standard deviation of the test scores.

From the above analyses and arguments, it appears that the probability of committing both errors (Type I and Type II) in an educational testing context might be reduced when the task is set at the medium level or middle range of difficulty (neither too easy nor too hard). Setting the probability of success too low (hard test or task) might increase the chance of making a Type II error (false positive). The possibility of making a Type I error (false negative) would be increased when the probability of success is put too high (easy test or task). Furthermore, in applied research, educational testing and therapeutic intervention settings, Type II errors might increase when using directional (one-tailed) hypotheses at $p < .01$ level of testing (Ary, Jacobs & Razavieh, 2002). The relative seriousness of a Type I or a Type II error is a judgmental decision made by individual researchers, clinicians and evaluators. In the present review, both of these two errors are considered serious. For instance, Type I error exaggerates the learner's ability. Lowering of the test difficulty raises spuriously the probability of success. Although they help improve interest and motivation by permitting experience of success, easy tasks may induce boredom, complacency and overconfidence. Conversely, a Type II error understates the learner's capacity. Hard tasks or tests unnecessarily lower the probability of success thereby increasing the likelihood of Type II error to occur. This can depress students' interest and motivation to learn. However, hard tests or low (highly stringent) levels of significance (e.g. $p = 0.001$) are useful in life-and-death studies, research and occupations (such as medical doctors and pilots) where chances of doing something wrong should greatly be minimized. Besides the false negative and false positive outcomes, there are also two other possibilities a test might produce, namely: the true negative and true positive. For example, the student with math difficulties who is correctly identified by the test as having math problems, is a true positive. The opposite or reverse is also true. The child without math difficulties who is correctly assessed/diagnosed by the test as having no mathematics problems, is a true negative. The ratio of correct or concordant identifications (true positives plus true

negatives) to incorrect hits (false positives plus false negatives) is sometimes used as an index of test reliability or as an indication of the efficiency of the assessment procedure (Bansell, 1986; William, 1991).

Caution on the use of optional questions in examinations

In some educational systems, examinations are curriculum oriented rather than measurement oriented. To cover more contents in a long syllabus a large number of essay questions (say 9) may be given to examinees in the subject (e.g. history) at one sitting who are then asked to select and answer any few questions (e.g. 3) they like/prefer. Optional questions should be very few (1 or 2) or totally discouraged because they give rise to many measurement problems (see Commonwealth Secretariat, 1973). Reducing the number of optional questions would ensure that examination candidates do the same test as far as possible and that the results are comparable for selection purposes. To achieve this, it would be better to include a compulsory question. Giving optional questions is a democratic process but causes a lot of measurement problems listed and briefly discussed below:

- For example, there would be 816 different combinations of 9 essay questions taken 3 at a time.
- Technically, students who chose to answer Items 1, 2, and 3 would have written a different exam from those who answered Items 4, 5, and 6 or those who answered Items 7, 8, and 9.
- Students' performance or achievement on different combinations of questions (different exams) cannot be reasonably compared because they will have written different questions.
- Questions on different and non-overlapping combinations will differ in difficulty, reliability, validity and bias.
- Examinees of different ability will choose and combine questions differently.
- Different combinations of questions cannot easily be scored in the same way by one or more markers.
- The skills tested in different combinations of the questions would not necessarily be the same.
- The higher the number of optional questions the more different combinations of questions there would be.

Types and quality of the assessments

Because most GenNEXT courses are interdisciplinary and multifaceted, they will be taught using a variety of instructional methods and assessed by a wide range of testing techniques. For example, some teaching may occur in the laboratory (e.g. experiments) and others in the field (e.g. action research), formal settings (e.g. classroom), community environment (e.g. outreach projects), and industrial sites (e.g. informal contexts suitable for work attachments, practicums, and internships). Similarly, the assessment instruments will also vary considerably. Both conventional and alternative assessment strategies will be employed depending largely on the type of course and skills that need to be measured. Assessment procedures might include: portfolios; problem-based assessments (PBA); performance-based assessments (also abbreviated as PBA); research projects; experiments; field activities; video or picture analyses; and informal assessments such as speech-making, interview, or role-play. From a teaching point of view, it would be highly unrealistic and self-defeating to view each of these assessment procedures in absolutistic/irrational terms such as "one-size-fits-all" or "all-but-nothing thinking style" when using them at UBD under the ongoing GenNEXT programs.

Regardless of the assessment techniques used, the Ministry of Education (SPN21 curriculum) and UBD (GenNEXT programs), will need to be wary of the quality of education. From the assessment standpoint, some of the factors that may impact the quality of education adversely include reliability, validity, and bias. Brunei will need to invest time, money and effort in conducting credible validation studies of the assessment strategies used in the country's entire education system. For example, content analyses of the SPN21 and GenNEXT examination instruments are needed to determine the skills measured by these assessments. The infrastructure for carrying out such research is already available in the Ministry of Education (Department of Examinations) and at UBD (Strategy and Quality Assurance Office). For the sake of fairness to students, there is nothing wrong about emphasizing quantity (multiplicity and diversification) in assessment procedures provided quantity is matched with quality. However, quantity alone without quality is always no good. Suggested in Table 2 below are examples of some procedures for assessing the reliability and validity of assessment instruments, both quantitative and qualitative (adopted from Mundia, 2001). This list is, however, not exhaustive but merely illustrative. In addition, bias detection techniques such as differential item functioning are not included in this list. Unfortunately, brevity considerations do not permit any full description of the reliability and validity detection techniques mentioned here. Although outside the objectives and beyond the scope of the present review, it is hoped that future reviews will address further these quality issues. Meanwhile, more attention and effort should perhaps be directed to validity concerns since any valid test would also be reliable (Rust & Golombok, 1989) but not necessarily vice versa.

Table 2

Procedures for assessing test reliability and validity

Dimensions	Quantitative tests	Qualitative tests
Reliability	Test-retest Parallel-forms Split-half Chronbach alpha ANOVA method KR20 KR21	Test-retest Inter-coder/inter-rater/inter-judge agreement % Intra-coder/intra-observer/intra-judge agreement % Inter-observer/intra-observer agreement % Member checking/verification Debriefing Cross-checking for erroneous entries
Validity	Content validity Concurrent validity Predictive validity Construct validity Discriminant validity Factorial validity Internal validity External validity Convergent validity	Multiple interviews Persistent observation Prolonged engagement Triangulation Ecological validity Social validity Comparison of videotapes with records Content validity Negative cases Rival explanations Audit trails External criticism Internal criticism Comparison of concept maps with records

Use of raw and standard scores in selection

Raw scores are essentially original scores students obtain from an assessment task. If the selection of students for admission, scholarships, prizes, or employment is based on a single test that does not have clearly distinct subsections/subparts/components, then the raw scores may be used for ranking and picking up the best students. However, if students were administered a battery of tests (e.g. of different subjects) or a single test with different components, then it might not be realistic and meaningful to use composite scores based on raw scores for selection purposes. Raw scores on different tests cannot be added up across the tests to get a composite score if they do not lie on the same scale. Summing them up in this way would, analogously, be like adding apples, oranges and bananas or adding British pounds, American dollars, and Euros together without first converting them to a common currency denomination. Both of these examples refer to adding things which are not the same. In mathematics or statistics you can, these days, add (subtract, divide, and multiply) any numbers using the computer for speed and accuracy and obtain any kind of answer (e.g. half or quarter of a person). The question to be asked here might be: is this answer meaningful or senseless? According to Lord, (1977; 1980) and Hambleton et al. (1991) the only time when raw scores from different components of a single test or from different subject tests can be added up across to obtain composite scores, is when they lie on the same scale (i.e. have the same or identical mean and standard deviation values). If not, then the raw scores should be standardized before adding them up. Standardizing scores means putting scores on the same scale or equating scores (to make them comparable/equivalent). Schools, colleges, and universities in Brunei, like elsewhere, frequently use scores for making selection decisions. Under the ongoing SPN21 and GenNEXT curriculum reforms, it is important that scores used for selection purposes are standardized (equated or equivalent in value) to avoid making erroneous selection decisions (e.g. awarding a degree classification, admission, scholarship, prize, or employment to a wrong recipient). This would be fair practice in using test scores that might help prevent possible litigations to courts of law following a wrong selection decision based on inaccurate test score analysis and interpretation. There are many ways raw test scores may be standardized or equated. The easiest examples are: Z-scores (mean = 0; standard deviation = 1); stanine scores (mean = 5, standard deviation = 2); and transformed or T scores (mean = 50; standard deviation = 10).

Role and use of external assessors

This is one way of obtaining the reliability and validity of the examinations and checking on unforeseen biases (all contributing to the quality of education). It is comparable to the process of inter-marker/inter-rater agreement reliability. There are many models of external assessor roles. In general, external assessors are quality specialists and are paid an honorarium or consultancy fee. They audit the course curriculum, test papers and marking keys before examinations are conducted. Feedback information on these is often given either verbally in form of a meeting with course instructors (internal assessors) or written report submitted to the respective Department or Faculty. Any anomalies detected prior to the administrations of the examinations are corrected. After the exams are done and completed, the same external assessors scrutinize, review and evaluate the performance of purposefully selected representative samples of the more able (high scoring) and less able (low scoring) students in their specialization areas or subjects. Students scoring in the middle of the range (neither too low nor too high) are usually ignored. Any anomalies detected in awarding marks and grades to students in these two extreme groups are discussed (verbally or in writing) and corrected by moderating the students' performance (adjusting marks and grades). Verbal and written feedback is provided to convey and explain the external assessor's actions. The main flaw under this system is, as pointed out above, not much attention is accorded to students performing in the middle of the range (average scorers). The external assessors are usually appointed from reputable universities in other countries. They are supposed to be highly qualified and experienced in assessing students in their areas of

specialization. The documents (syllabi, test papers, and marking keys) to be analyzed are either sent to the assessors (cheaper mode) or the assessors are brought to the university (a more expensive option as it involves paying airfare, accommodation, and meal costs in addition to an honorarium or consultancy fee). The system has both pros and cons but no research has, so far, been conducted to determine its efficiency and effectiveness in Brunei tertiary institutions. Under the ongoing GenNEXT reforms at UBD, some courses (few) have external assessors. UBD might wish to review the system of external assessors to make it more effective. This might require appointing external assessors for all the courses, if necessary. The system is quite important as it contributes to improving the quality of education and might help gain international recognition and ranking of the university and its graduates. Furthermore, the system includes the external assessment of graduate students' research (masters and doctoral theses and dissertations). Moreover, if extreme care is taken to appoint only professionally qualified external assessors in some areas/fields, the external assessments might form the basis for securing the accreditation and recognition of UBD courses and credentials with international professional associations or organizations (e.g. in psychology, counseling, medicine, and engineering).

Conclusion

As can be deduced from this review, there are many other assessment-related issues that need to be addressed in Brunei secondary school and higher education institutions. One of them is computer-assisted assessment. This rapidly expanding area has potential to impact the assessments positively at both the secondary and tertiary levels of education in Brunei particularly with the exceptional student populations. The problem is that it requires considerable user technical skills in both assessors and students. In addition, the high cost of computers and relevant software programs may be prohibitive and render the procedure not feasible for large-scale application. This leads to the second major issue. If assessments were to contribute extensively in improving the quality of education, then UBD and the Ministry of Education need to consider giving advanced training in educational testing, measurement and evaluation to selected lecturers and school teachers, respectively. The selected recipients of advanced training could, through the multiplier effect, pass on the mathematical, statistical, technical, and qualitative skills to colleagues in their school, department, or faculty via in-service CPD workshops. Advanced kills in educational assessment and evaluation (see Lord, 1977, 1980; Hambleton, 1982; Mundia, 2000) are also extremely useful to teachers and lecturers in doing high-quality research. Overall, the current and ongoing curriculum and assessment reforms in the Brunei education system are likely to succeed given the commitment and financial support from the country, institutions and the public. To gain additional insights on the issue, further mixed-methods research on assessments is recommended at all levels of education in Brunei. Findings from such research might also be useful elsewhere to other small universities in small nations.

Author Biography

LAWRENCE MUNDIA teaches educational psychology in the Hassanal Bolkiah Institute of Education at the University of Brunei Darussalam. His main research interests are in psychological assessment, educational testing, special education, and school guidance and counselling. Recently, students with mental health problems have received more research consideration, attention and priority. He is keen on researching female students' mental health problems in tertiary institutions.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

Ary, D., Jacobs, L. C., & Razavieh, A. (2002). *Introduction to research in education.* Belmont, CA: Wadsworth/Thomson Learning.

Bansell, R. B. (1986). *A practical guide to conducting empirical research.* London: Harper and Row.

Bartram, D. (1995). The Development of Standards for the Use of Psychological Tests in Occupational Settings: The Competence Approach. *The Psychologist*, May, 219-223.

Bartram, D. (1996). Test Qualifications and Test Use in the UK: The Competence Approach. *European Journal of Psychological Assessment,* 12, 62-71.

Boyle, J., & Fisher, S. (2008). *Educational testing: A competence-based approach.* London: Wiley-Blackwell.

Clark, C. (2009). SPN21 briefing at St. James school. *Borneo Bulletin*. Saturday 27 June 2009, page 11.

Commonwealth Secretariat. (1973). *Education in the commonwealth, No. 18: Public examinations.* Report of the Commonwealth planning seminar held in Accra, Ghana, 1973. London: Commonwealth Secretariat.

Eyde, L. D., Moreland, K. L., & Robertson, G. J. (1988). *Test user qualifications: A data-based approach to promoting good test use.* Report for the Test User Qualifications Working Group. Washington DC: American Psychological Association.

Eyde, L. D., Robertson, G. J., & Krug, S. E. (1993). *Responsible test use: Case studies for assessing human behavior*. Washington DC: American Psychological Association.

Fremer, J., Diamond, E. E., & Camara, W. J. (1989). Developing a code of fair testing practices in education. *American Psychologist*, 44, 1062-1067.

Hambleton, R. K. (1982). Advances in criterion-referenced technology. In C. R. Reynolds and T. B. Gutkins, The handbook of school psychology (pp. 351-379). New York, NY: John Wiley & Sons.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage Publications.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229-244.

Joint Committee on Testing Practices. (1988). *Code of fair testing practices in education*. Washington DC: Joint Committee on Testing Practices.

Joint Committee on Testing Practices. (2000). *Rights and responsibilities of test takers: Guidelines and expectations*. Washington DC: Joint Committee on Testing Practices.

Joint Committee on Testing Practices. (2004). *Rights and Responsibilities of test takers: Guidelines and expectations*. Washington DC: Joint Committee on Testing Practices.

Kuder, G., & Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrica*, 2, 151-156.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement*, 14, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Ministry of Education. (1997). *Special education policy guidelines.* Bandar Seri Begawan: Unit Pendidikan Khas.

Ministry of Education. (1998). *Handbook on special education for primary school headmasters: Outline of policies and procedures dealing with special education*. Bandar Seri Begawan: Unit Pendidikan Khas.

Ministry of Education. (2007). *Proposed SPN21 curriculum: Draft.* Bandar Seri Begawan: Curriculum Development Division, Ministry of Education.

Moreland, K. L., Eyde, L. D., Robertson, G. J., Primoff, E. S.,Jac & Most, R. B. (1995). Assessment of Test User Qualifications: A Research-Based Measurement Procedure. *American Psychologist*, 50, 14-23.

Mundia, L. (1999). *Practical guide to introductory measurement and evaluation in education and psychology.* Madang, Papua New Guinea: Kristen Press.

Mundia, L. (2000). *Basic statistics for introductory measurement and evaluation in education and psychology*. New Delhi: UBS Publishers' Distributors.

Mundia, L. (2001). *Fundamentals of quantitative and qualitative research in education and psychology*. Matsapha, Swaziland: Jubilee Printing & Publishing.

Murray, D. (1996). Forget Charity? We have a right to fair assessment: accommodating students with disabilities need not compromise assessment standards. Paper presented at the conference on "Partnerships on the Assessment of Student Achievement" held in Auckland, New Zealand, September 22 – 28.

William, P. (1991). *The special education handbook: An introductory reference*. Milton

Keynes: Open University Press.

Ory, J. C., & Ryan, K. E. (1993). *Tips for improving testing and grading*. Newbury Park, CA: Sage Publications.

Rust, J., & Golombok, S. (1989). *Modern psychometrics: The science of psychological assessment.* London: Routledge.

Somerset. H. C. A. (1987). *Examinations reform: The Kenya experience.* Report No. 64. A Report Prepared for the World Bank. Sussex: IDS.

UBD Brochure. (2012). *Universiti Brunei Darussalam – Our National University*. Bandar Seri Begawan: Universiti Brunei Darussalam. Available online at: http://www.ubd.edu.bn