CHAPTER THREE

LITERATURE REVIEW AND CONCEPTUAL FRAMEWORK

3.1 INTRODUCTION

An extensive literature search was undertaken of primary and secondary sources, including books, paper-based and electronic journals, databases, and conference proceedings and conference papers. The literature review began with an internet search and proceeded through *ERIC* and the Universities of Pretoria and Botswana databases. Key words used in the search were assessment, assessment for learning, authentic assessment, performance assessment, constructivism, pragmatism, validity of assessment, reliability of assessment, formative assessment, and quality assurance in performance assessment.

The literature review was guided by two main research questions, which sought to find out:

1. *How valid and reliable are the performance assessment processes in Botswana?*

2. *How can quality assurance processes be developed in order to produce valid and reliable marks for BGCSE Agriculture performance assessment?*

The next Section, 3.2, gives a brief background of the origin of performance assessment, followed by conditions for performance assessment in Section 3.3. Section 3.4 outlines quality assurance of performance assessment internationally. Issues in performance assessment are outlined in Section 3.5 after which conditions of performance assessment in Botswana are delineated in Section 3.6. The discussion focuses on how validity and reliability of performance are ensured internationally in Section 3.7. The conceptual framework of the study is delineated in Section 3.8. The conclusion/synthesis of literature review is presented in Section 3.9.

3.2 THE ORIGINS OF PERFORMANCE ASSESSMENT

Performance assessment has been in existence for a long time. Madaus and O'Dwyer (1999) trace the origins of testing to China where it was applied to different disciplines, such as Letters, Law, History, Rituals and Classical Study. It was applied to Education around the early eighteenth century (Morris, cited in Johnson et al., 2009), mainly as oral examinations, and replaced by essay examinations around 1845. Airasian and Russell (2008) posit that "performance assessment has been used extensively in classrooms for as long as there have been classrooms" p. 205.

During those early years, the judgement of the examinees' performance was mainly qualitative (Hoskin, 1979 Johnson et al., 2009), which introduced the problems of subjectivity and partiality, especially when high-stakes decisions were made. Apart from subjectivity, other problems are such as unreliability of scoring the essay exams were common (Starch & Elliot, cited in Johnson et al., 2009). This did not go unnoticed as the search for better ways of assessing commenced, leading to the invention of multiple-choice tests in 1915 by Frederick Kelly. The invention of multiple-choice testing prefaced the development of standardised, norm-referenced tests (Madaus & O'Dwyer, 1999) and marginalised performance assessment.

During the 1980s and 1990s, the assessment community witnessed the resurgence of performance assessment in education (Johnson et al., 2009). As testimony to this, Stiggins (1995) titled his textbook on performance assessment, *An Old Friend Rediscovered,* in reference to performance assessment. Today performance assessment plays an important role in examinees' lives, as assessment bodies and commercial examination providers embrace the incorporation of performance assessment marks in certification (Berry, 2008). Clauser, Harik and Margolis (2006) write that performance assessment has increasingly been used as part of high-stakes testing programmes during the past decade, because in some situations it is inevitable.

3.3 CONDITIONS FOR PERFORMANCE ASSESSMENT

The distinct characteristics of performance assessment which warrant its implementation are discussed in the subsequent subsections. Performance assessment requires different conditions from those of a paper-and-pencil test. These conditions are such as assessment for learning, assessment enhancing abstract and creative thinking, assessment of authentic tasks, catering for students' cognitive differential development, and complex content which encourages critical thinking.

Assessment for learning

Research in assessment has traditionally been concerned with studies of the validity and reliability of the externally designed and administered tests and examinations, which were held in high esteem (Black 1993; Harlen, 1994; Popham, 2005). Those for performance assessment were and are still not given much attention. However, the purpose of learning and assessment has since changed from selection, guidance, and prediction of future performance (Stiggins, 2002) to accountability of the school and the education system as a whole (Airasian & Abrams, 2002). This heralded the switch from a testing culture to assessment which focused on encompassing evaluation of learning progress by the learner (Gasemann, 1993), with the provision of useful feedback to learners being the hallmark of assessment for learning (Nitko & Brookhart, 2007; Thorndike & Thorndike-Christ, 2010).

According to Assessment Research Group (ARG) (2002), Assessment for Learning as opposed to Assessment of Learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there. Assessment for learning gives the kind of challenges, diversity and flexibility that make assessment more realistic and educative rather than testing which simply audits learning (Wiggins, 1998):

> If we want to improve education to advance our standard of living, we must do away with testing and embrace assessment. Testing is characterized by secrecy and security. When tests are administered, rigid rules are followed (p. 14).

Nitko and Brookhart (2007) posit that educative assessment is authentic and involves showing students by doing, which motivates them to perform better than the awarding of marks. Such assessment has proved to be a powerful school improvement tool, as well as raising students' achievement to unprecedented levels. This led to Neill and Medina (1992) and later Wiggins (1998) to advocate for the abolishment of simultaneous group administration of paper-and-pencil tests which do not take into consideration students' readiness. A decade later, Lissitz & Schafer (2002) supported the reduction of emphasis on large-scale testing. Performance assessment allows students to demonstrate in a variety of ways their understanding, using knowledge and skills learnt from different areas. Diez (2002) concurs with Nitko and Brookhart and Wiggins, but proposes the balancing of classroom-embedded assessment with high-stakes measures.

Research summarised by Black and William (1998) shows that student self-assessment skills, learned and applied as part of formative assessment, enhances student achievement. In 1994, the Averno faculty (McMillan, 2000, p. 211) identified ten elements which when applied in assessment for learning could enhance students' developmental learning processes:

1. Explicit outcomes - clear picture of expectations of candidates' knowledge and performance through the outcomes that guide course and programme development through the dozens of assessment that candidates complete.

2. Performance - assessment of candidates' performance of what they can do with what they know.

3. Public, explicit criteria - criteria which describe the expected quality of performance and must be met.

4. Feedback - feedback from assessors or peers about weaknesses and strengths and how to improve.

5. Self-assessment - assessing oneself, which helps one to become his/her own coach and critic.

6. Multiplicity - assessing more than once using a variety of assessment methods and contexts over time.

7. Externality - trying out in real world situations or bringing others to help assess so as to avoid subjectivity.

8. Developmental nature - assessment which fits the candidates' developmental ability and knowledge.

9. Cumulative nature - assessment which is continuous to give a clear picture of knowledge and skill of students. Students grow over time, therefore are bound to show an improvement.

10. Expansive nature - assessments are developed to elicit from the candidate the most advanced performance of which each is capable.

However, there is no evidence suggesting adoption of these in Botswana as evidence by Mogapi & Yandila (2001) and Yandila, Komane & Moganane (2003) presented in Section 3.6 suggests the contrary.

Encouraging complex, abstract and creative thinking

There is no single correct answer to real-world problems, and standardised testing follows rigid regulations which could lead to failure by students to engage in demanding creative tasks (ARG, 2006; Shepherd 2000, 2008). In most cases, standardised paper-and-pencil tests are of low-order measuring how much learning has taken place with little regard to context, creativity and processes. Assessment as a social construction should consider students' social background and students' prior knowledge. For example, a *pest* in one region could be an extremely valuable creature in another region. Setting a multiple-choice question on such a "pest" could disadvantage students from a region where it is not regarded as a pest, so a question on *pest* should rather be set to encourage creative and thoughtful application and meaningful use of knowledge to solve problems caused by pests.

Assessment in performance tasks provides the opportunity to assess thinking processes that the students undergo to construct their responses (Airasian, 2005). The assessor observes the students performing a task to find an answer to a problem. The teacher then marks the students' every step taken and guides the student to the right procedure whenever the student deviates. The intention is not to assess the ultimate answer, but how the answer was arrived at (Airasian & Russell, 2008; McMillan, 2004). The former is the goal of selection type of assessment. It assumes that when the student gets an item correct, the student must have followed the correct process, but there is no direct evidence to support the assumptions. As such, performance assessment helps to gauge what pupils can do as opposed to selection tests that assess what pupils know (Neill, & Medina, 1992).

Authentic Assessment

Performance assessment should resemble the activities taking place in the real world (McMillan, 2000). According to Diez (2002), Rennert-Ariev (2005) and Ryan (2006), performance tasks address demanding tasks which normally span longer periods, hence requiring students to use many different skills and abilities. For example, students growing *a crop* spend at least four months managing it, and during that period they are engaged in a number of management activities. In carrying out these activities students are required to apply knowledge and skills acquired from different areas, including affective skills.

This kind of learning is authentic in nature (Johnson et al., 2009; Nitko & Brookhart, 2007) because students perform in the context of the real-world situations in which the skills are to be applied. They are involved in doing rather than just knowing how to do it or simply know it. Authentic skills are not fixed, hence they cannot be assumed to be conducted under standardised conditions or manifest themselves always in the same way at any time across contexts (Airasian, 2005; Popham, 2005). This calls for various formats and methods to be used for assessing students (Stiggins, 1997).

Wiggins (1998) developed a set of six standards for judging the degree of authenticity in assessment:

a) *Is realistic*: the task replicates the ways in which a person's knowledge and abilities are tested in real world situations.

b) *Requires judgement and innovations:* the student has to use knowledge and skills wisely and effectively to solve unstructured problems, and the solution involves more than following a set of routine procedures or plugging in of knowledge.

c) *Asking the student to do:* the student has to carry out the exploration and work within the discipline of the subject area, rather than restating what is already known or what was taught.

d) *Replicates or simulates the context in which adults are tested in the workplace, in civic life, and in personal life:* contexts involve specific situations that have particular constraints, purposes, and audiences. Students need to experience what it is like to do tasks in the workplace and other real life contexts.

e) *Assesses the student's ability to efficiently and effectively use a repertoire of knowledge and skills to negotiate a complex task:* students should be required to integrate all knowledge and skills needed, rather than to demonstrate competence of isolated knowledge and skills.

f) *Allows appropriate opportunities to rehearse, practice, consult resources, and get feedback on and refine performances and products:* rather than relying on secure tests as an audit of performance, learning should be focused through cycles of performance – feedback – revision - performance, on the production of known high-quality products and standards, and learning in context (p. 22-24).

However, in some situations, conducting authentic performance assessment is unattainable, such as when performance is complicated and equipment is expensive, or puts other people's lives in jeopardy. For example, the application of chemicals to control pests by students under the age of sixteen is not legally allowed. In such situations, simulations could be an alternative (McMillan, 2004) to serve as an intermediate step to performance that involves a higher degree of realism.

Catering for different developmental rates

Students have been found to develop intellectually at different rates, depending on their background, experiences and learning styles (Neill, & Medina, 1992). Since learning is related to intellectual development, it should follow therefore that learning should also be differentiated to cater for individual differences. The multidimensionality of students' development requires learning to be based on theories that encompass dimensions of cognitive, psychomotor and affective skills (Nitko & Brookhart, 2007). Such learning provides an opportunity to students who do poorly on the cognitive dimension to show their achievement in performance assessment (Airasian, 2005). Assessing students in multiple ways provides the opportunity for students to engage in performance assessment, which renders equal opportunity to be assessed in all domains of development. Research has found that individuals exhibit different ways of knowledge and problem solving that reflects different styles, not different abilities, yet standardized paper-and-pencil tests assume that all individuals perceive information and solve problems in the same style (Neill & Medina, 1992).

Covering complex content

Performance assessment covers in-depth content of knowledge and skills (Johnson et al., 2009), the coverage of which comes with the problem of relatively few tasks being used as compared to other formats of assessment, resulting in scores suffering from external validity (Lane & Stone, 2006). Airasian (2005) advises that this could be overcome by properly created performance assessment tasks which sample a wide range of abilities to be applied by the student in solving complex problems. Wiggins (1998) cautioned against the use of low-order thinking skills as the solution to external validity. Performance assessment allows for the assessment of students' complex processes, as well as their product. Assessment of processes is crucial for the accomplishment of quality products and activities that have momentary evidence that cannot be formatively assessed by paper-based tests (Black, 1995; William & Black, 1996). These thought-provoking tasks have multiple solutions to allow students to construct their own meaning fostering development of thinking in varied styles.

In conclusion, performance assessment tasks should be essential, drawing from the core curriculum and representing a "bigger idea". The tasks should be authentic, using processes appropriate to the discipline, and students should value the outcome of the tasks which in turn lead to problems that require them to draw from deeper faculties in solving, rather than replicating known procedures. Quality assurance processes should be in place to validate performance assessment marks. Having looked at the conditions of performance assessment, the discussion now focuses on international examples of quality assurance.

3.4 QUALITY ASSURANCE OF PERFORMANCE ASSESSMENT INTERNATIONALLY

Performance assessment has become a necessary undertaking for many examination boards, with the main focus on quality assurance (Khoo & Idrus, 2004; Maughan, 2004), defined by Oakland (1993, p. 13) as "broadly the preventing of quality problems through planned and systematic activities (including documentation)". Performance assessment is premised on entrenching quality in the system and continual auditing and reviewing (Walklin, 1992; Doherty, 1994).

Although school-based performance assessment has been criticised for its lack of reliability in particular (Chong, 2009), it is necessary to maintain the right balance between teachers' professional judgment and national testing for national assessment systems to be comprehensive, rigorous and meaningful, while at the same time improving teaching and learning (Queensland Studies Authority, 2009; Pellegrino, Chudowsky & Glaser, 2001). All the same, there is no single solution in achieving this, and different countries employ different strategies or the same strategies applied differently, as per the dictate of their contextual factors (Broadfoot, 1994; Maxwell, 2004; Raivoce & Pongi, 2000). Despite the combinations and permutations possible, the first step in ensuring quality in performance assessment is to embed quality into the processes (Campbell & Rosznyai, 2002; Richard, 1993), which includes but not limited to teacher development of tasks, training to assess; resources provision; leadership commitment; development of learner/support materials, moderation, authentication, internal monitoring, external monitoring and supervision, multi-rating, and school approval, (Chong, 2009; Khoo & Idrus, 2004).

Training of teachers to acquire the appropriate expertise is essential (Broadfoot, 1994), as the public has confidence in trained teachers to conduct assessment professionally and ethically (Maxwell, 2004). Germany and Australia are well known for emphasising professional development of teachers to assess (Broadfoot, 1994; Queensland Studies Authority, 1998), because their assessment procedures are largely the responsibility of teachers, even for certification and selection purposes, with minimal external intervention or moderation (Gasemann, 1993).

Teachers in Germany and Australia develop their own good quality assessment tasks and procedures, therefore assessment marks are not aggregated by a mathematical formula to produce an overall result. Rather, the result involves an interpretation of the final product of the student's work by a judgement of the standard it demonstrates when compared to a set of grade descriptors (Mercurio, 2008; Maxwell, 2004). Teacher training is not a sufficient condition but high management commitment to quality provision can enhance their performance (Burdett & Johnson, 2009). The management should be well-versed in assessment practices to manage assessment process and support teachers (Bennett & Taylor, 2004; Calvo-Mora, Antonio Leal & Roldan, 2006; Wild & Ramaswamy, 2008).

The importance of moderation in ensuring comparable outcomes and improving teachers' assessment capabilities through applying agreed standards consistently by the individuals involved is very important (Queensland Studies Authority, 2009). According to Klenowski and Wyatt-Smith (2008): "Moderation can no longer be considered an optional extra and requires system-level support especially if, as intended, the standards are linked to system-wide efforts to improve student learning." (p. 1). Jordan and McDonald (2008), Masters and McBryde (1994), and Stanley and Tognolini (2008) note that the inter-rater reliability of the moderation system practiced in Queensland, in which teachers and schools were accountable for the assessment and reporting of student achievement, surpassed that of many external examination regimes. Furthermore, Bennett and Taylor (2004) assert that a system of moderation of teachers' judgments through professional collaboration benefited teaching and learning as well as assessment. Such a moderation procedure has more than a quality assurance function.

Moderation as a social ratification of teachers' assessment (Radnor & Shaw, 1995) is directed towards ensuring that quality assessment standards have been applied consistently. Moderation directed at ensuring quality is the most commonly applied form of moderation in Britain, New Zealand, Malta, Kenya, and Australia, to mention only a few countries (Boustead, 2008; Broadfoot, 1994; Harlen, 1994; Maxwell, 2004; Onyango & Ndege, 2007; Raivoce & Pongi, 2000). However, moderation employing a variety of methods that combine both quality assurance and quality control procedures has been found to yield better results (Berry, 2008; Keightley & Coleman, 2002; Queensland Studies Authority, 2009; Maxwell, 2004; Raffan, 2000; Raivoce & Pongi, 2000). For example, members of the markers' panels of different subjects visit the schools to moderate the candidates' coursework (Grima & Ventura, 2000; Queensland Studies Authority, 2008).

In SPBEA, a two-level moderation procedure is conducted, using students' samples to account for any differences between schools within a country and between countries, while statistical moderation is carried out on Teacher Designed Tasks. In the USA, the assessment is validated and calibrated using three models, namely: i) a national exam for a sample of students in each grade level, used to verify standards assessed by regional or local examinations; ii) some element of national exam with local exam; or iii) marking visiting teams cross-moderate between schools (Broadfoot, 1994).

While a few countries still moderately employ statistical moderation, such as Sweden, Hong Kong, South Africa, and to some extent New Zealand, (Berry, 2008; Broadfoot, 1994; Lennox, 2000; Singh, 2004), majority have abandoned its use (Broadfoot, 1994; Harlen, 1994; Maxwell, 2004; Radnor and Shaw, 1995; Raivoce & Pongi, 2000), in favour of embracing moderation processes that ensure quality (Boustead, 2008; Keightley & Coleman, 2002; Lennox, 2000), Critics of statistical moderation argue that this constitutes a typical misuse of statistical tools:

> the choice of a theory examination as reference standard for the moderation of practical grades is not beyond criticism. This inadequate and lazy practice is no longer allowed and statistics are now used in support of more relevant techniques of moderation, (Kempa, 1986, p. 85).

> statistical moderation, which often uses as its external reference point a written public examination, can stifle innovation in the classroom, and, in particular, can whittle away the professional skills of the teacher to design the assessment and make appropriate judgments (Mercurio, 2008, p. 9).

School approval or accreditation is another way of ensuring quality in performance assessment (Council for Higher Education Accreditation [CHEA], 2002), as it involves an evaluation of the capacity of a school to enter students for the board's qualifications, and provides an opportunity for schools to improve their management, learning and assessment processes. Before the school is allowed to conduct performance assessment, an audit of its capabilities to successfully implement performance assessment is conducted (CHEA, 2002; Colbeck, Caffrey, Donald, Lattuca, Reason, Strauss, Terenzini, Volkweinm, & Reindl, 2000; Jones, 2002). Officers from the examination board visit schools during the year in order to evaluate the physical and human resources available to carry out the coursework as specified in the syllabus, and the type and standard of the coursework.

Officers also observe and evaluate the assessment methods and procedures (Grima & Ventura, 2000). Participating schools are required to submit an assessment programme, clearly indicating what they intend to do, as well as the various assessment tasks that make up the programme for each subject that has a performance assessment component (Keightley & Coleman, 2002; Raivoce & Pongi, 2000). For example, SPBEA has to check for compliance with such factors as prescribed requirements, appropriate standards, and timeframe (Raivoce & Pongi, 2000) before any programme is approved for implementation.

After the school has been given permission to implement performance assessment, monitoring is carried out regularly to ensure that assessment is carried out in a satisfactory manner and that the school complies with the specified standards. Schools are required to maintain quality assurance systems to continue to carry out the Board's qualifications, and monitoring of adherence to standards is done both internally and externally. Selected schools are visited each year to verify that internal assessment programmes are being followed and to assist teachers in the delivery of the learning programmes (Keightley & Coleman, 2002).

One fundamental aspect of ensuring quality in performance assessment is by developing task frames to guide the development of tasks (Keightley & Coleman, 2002). Task frames are summary prose statements which detail the types and range of performance, representing different levels stipulated in the national curriculum. In Germany, France and Australia, performance tasks are developed by teachers themselves, after undergoing vigorous training (Broadfoot, 1994; Keightley, 2002), whereas in Sweden and Singapore task development is highly centralised (Chong 2009; Maughan, 2004).

The SPBEA, for example, uses three approaches to developing tasks: i) centrally developed tasks known as Common Assessment Tasks; ii) Teacher Designed Tasks (TDTs) developed by individual teachers in school; and iii) Common Assessment Frame tasks (CAFs), all determined by SPBEA but the tasks of which are developed by teachers (Ravoice & Pongi, 2000) then evaluated according to the achievement standards pre-defined to judge students' ability to meet the expected level of assessment (ARG, 2006; Keightley & Coleman, 2002). The details of developing tasks shall be discussed in detail in Section 6.3.

Scoring of performance assessment presents a problem as it sometimes involves judgement. The subjectivity of performance assessment is greatly reduced if multiple rating is used (Airasian, 2005; Airasian & Russell, 2008; Thorndike & Thorndike-Christ, 2010), with research having proved that the same test could be scored differently by different teachers and that even the same teacher could score responses differently at different times (Rennert-Ariev, 2005). According to Rudner and Boston (1994), multiple ratings improve reliability, in as much as multiple test items can improve the reliability of standardised tests. Further improvement of the reliability can be made by the use of criteria in scoring (Nitko, 2004). Throughout the process of scoring, scorers recalibration of raters through refresher practice sessions (Johnson et al., 2009), to avoid 'rater drift' should be done (Becker & Pomplun, 2006).

It is evident from the above discussion that the majority of countries were moving towards embracing performance assessment, to compliment standardised testing. While in some countries performance assessment tasks were centrally developed, they were solely the responsibility of the teacher in other countries. Similarly, the magnitude of quality insurance fell along the continuum of professional development of teachers to moderation. Forms of moderation applied ranged from the controversial statistical moderation to the consultative visiting moderation (Radnor & Shaw, 1995). In other countries, the implementation of performance assessment is fully developed to an extent of completely depending on the assessment conducted by teachers for certification, without any moderation or external intervention.

Though performance assessment allows the teacher to provide the information about what the student can do, its conduct remains problematic. The general public has not yet accepted it as a formal way of impartially assessing students, and even professional teachers seem not to understand their role adequately (Chong, 2009). This is because teachers' pedagogical training does not normally emphasise performance assessment.

Given the above factors that can enhance quality performance assessment the following section discusses issues in performance assessment that hinders its effective conduct.

3.5 ISSUES IN PERFORMANCE ASSESSMENT

The past two decades have witnessed a global trend towards performance assessment (Airasian & Russell, 2008; Abraham, 2008; Berry, 2008; Crooks, 2004; Harlen, 1994; Harlen, 2006; Maughan, 2004; Maxwell, 2004; Pongi, 2004; Raffan, 2000; Raivoce & Pongi, 2000). For example, over 40 states of the USA had adopted some form of performance assessment by 2000, (Patchen, 2004), while in the United Kingdom (James 1994) every school curriculum subject has introduced performance assessment as a component in the past 20 years (Raffan, 2000; Berry, 2008). In the Hong Kong education system, performance assessment was introduced as an important aspect of the assessment reforms (Hamp-Lyons, 2009), while in SPBEA it was introduced as a way of striving to provide quality and timely service to its clients (Pongi, 2004). In India, the 2005 National Curriculum Framework proposed a shift from traditional assessment based on behaviourism to constructivist approaches (Kapur, 2008).

However, performance assessment is not without problems (Chong, 2009). A discussion of the issues that arise in the use of performance assessment follows, looking at the problems, the debates, and how specific countries have dealt with these, as well as issues that have not yet been resolved. Major problems in performance assessment will now be examined in turn.

Development of tasks

Variation in the demand of tasks or opportunity provided by the tasks undertaken by students is one issue that is problematic in performance assessment (Department of Education 2001, cited in Singh, 2004). The inability by teachers to develop appropriate materials for assessment purposes, consistent with the relevant national curriculum (Kanjee & Sayed, 2008), is due to lack of training (Chong, 2009; Maxwell, 2004; Nenty, Odili and Munene-Kabanya, 2008; Stiggins, 2000). Developed countries are making progress towards entrenching quality in performance assessment among teachers (Broadfoot, 1994; Maxwell, 2004). This is because teachers' technical competence to assess invariably facilitates the interpretation of performance criteria. To prevent the intrusion of irrelevant contextual information in making judgements, marking schemes are to be understood and applied in the same way.

Maxwell (2004) asserts that if teachers are properly trained and given enough support resources, they can design and develop sound assessments, which can then be used to:

i) determine what a student has learnt and what s/he still needs to learn,

ii) help each student learn and use knowledge well,

iii) determine how well the teacher applied an instructional process, and

iv) provide information to students, teachers, and parents (Mamary, 2007, p. 188).

Provision of resources

Implementing performance assessment on a large scale requires massive resources, which are costly (Tindal & Haladyna, 2002), but consequential gains to the learner are immeasurable. Doty (1996) suggests that costs of implementing performance assessment can be significantly reduced by identifying and controlling expenses through budgeting, measuring, and analysis, to achieve higher quality education at lower cost. Since performance assessments are perceived to be expensive, as in the case of portfolio which is developed over a period of a year with many students in a class (Mills, 1996; Johnson et al., 2009; Nitko & Russell, 2007), limited resources and time are often directed towards less expensive standardised testing (Stiggins, 1997). Pellegrino, Chudowsky & Glaser (2001) called for the balance of mandates and resources to be shifted from an emphasis on external forms of assessment to an increased emphasis on classroom formative assessment, given that well-resourced schools tend to perform better (Howie & Plomp, 2001).

Teacher Workload

Performance assessment as a student-centred approach requires more time for individualised instruction (Fung et al., 1998) and recording of the student achievement and progress. As a result, some teachers view school-based assessment as an extra workload imposed by an external institution (Keightley & Coleman, 2002; Torrance, 1995), which should be paid for particularly when done for summative purposes (Grima & Ventura, 2000). Because of the work involved, teachers prefer externally set practical examinations to school-based assessment (Raffan, 2000), despite the well-documented validity evidence of the later. Teachers who then engage in school-based performance assessment resort to inflating students' marks under the pretext of time constraint (Raivoce & Pongi, (2000).

It is generally accepted that class size and workload are related, however less clear is whether class size has any effect on achievement. Mixed and inconclusive findings have been reported about the effect of class size, for example Finn and Achilles (1990), Hoxby (2000), Milesi and Gamoran (2006), Nye, Hedges and Konstantopoulos (2002), Pong and Pallas (2001), found little or no gain in small class sizes. On the other hand, Angrist and Lavy (1999), Knostantopolous (2008), Knostantopolous and Chung (2009) contend that small class sizes yield positive results, particularly in developed countries. Finn et al. (2003), Jones (2006), Miller, Sen and Malley (2007) identified gains for small class sizes to be : (i) more participation, engagement and identification; (ii) more teacher time per student; and (iii) more time for individualised assessment and increased time on task.

However, class sizes were found to be large in African and Asian schools (Bery, 2008). For example, in Malawi, class size in 1994 at primary level was 100 (Nowa-Phiri, 2000), while in South Africa grade 8 average Mathematics class size was 46 students (Howie & Plomp, 2001). Large class size was found to be an impediment to implementing authentic assessment (Howie, 2006). Howie and Plomp (2003) found that class size and work load affected students' performance in mathematics. If there are too many students in the classroom, the teacher's assessment focus tends to be on class, or perhaps the small group, rather than the individual student.

Low weightage

The contribution of performance assessment towards final grade varies significantly from country to country, depending on the development of the structures in place as well as the confidence the public has in performance assessment outcomes (Chong, 2009). Nitko (1995) proposed three models for combining performance assessment results with National examination results:

> Model one: using performance assessment only at school level but not counting them toward certification
>
> Model two: count performance assessment toward certification or selection using a compensatory model (e.g. regression weighting)

> Model three: count performance assessment toward certification or selection but fix the percentage weight (e.g. 40% or 60% of the total performance assessment: (a) count only the last few years, (b) count all years, or (c) count all years but weigh earlier years less than later years. (p. 5)

A number of countries have reported varying contributions of performance assessment, even between subjects within a country. In England and Wales, for example, significant elements of teacher assessed coursework and practical work was weighted between 20% and 100 %, depending on the subject and syllabus followed (Torrance, 1995). The weight of performance assessment component for PSSC countries ranged from 40% to 100%. (Raivoce & Pongi, 2000; Ventura & Murphy, 1998), while in Germany, performance assessment as the responsibility of teachers contributed 100%, with minimal external intervention (Broadfoot, 1994; Gasemann 1993).

In the UK, performance assessment in the non-core subjects contributes 100% of the final statutory assessment (end of Key Stages 3 and Stage 14), while at the GCSE level varies. For example, Biology's coursework contributes 20% (Maughan, 2004). In Australia, performance assessment's contribution varies from one province to another, between 50% and 100% (Keightley & Coleman, 2002). For example, in Queensland it contributes 100%, hence there has been no standardised public examinations for over 35 years (Maxwell, 2004).

In Kenya, CA was implemented in only three subjects and ranged from 10% to 25% (Noor, 2008), while in Namibia, summative CA contributed between 30% and 50% to the end-of-course grade (van der Merwe, 2000). Njabili (1987) reported CA's contribution to be 50% across the board in Tanzania. In South Africa, CA contributed 25% to the final Grade 12 examination grade (Kanjee & Sayed, 2008; Singh, 2004; Van der Berg & Shepherd, 2010).

Teacher training

It has been discussed under 'task development' that assessment should be all-encompassing. Research has revealed that teachers lack skills to develop tasks that can recognise full range of achievements of all students, despite the fact that they are the appropriate assessors of what is inaccessible to the external examination (Pellegrino, Chudowsky & Glaser, 2001; Tindal & Haladyna, 2002; Wiggins, 1998). Nevertheless, building the capacity and competency of teachers to carry out assessment in the classroom effectively and consistently is a challenging task (Chong, 2009). Due to inadequate training, teachers are not prepared to assess their pupils, especially on performance tasks (Kellaghan & Greaney, 2003).

Howie (2006) pointed out that one of the reasons teachers in South Africa cannot implement performance assessment successfully is because they have had insufficient training in assessment. Stiggins (2002) noted that about only one quarter of states in America require that pre-service teachers take an assessment course, but only three states require competence in assessment as a requirement or condition of being licensed as a principal, while no state certifies that competence.

Lack of training in assessment results in teachers deemphasising or neglecting untested materials (Tindal & Haladyna, 2002), as well as testing students on trivial outcomes which seek to find out whether the child knows, understands or can perform predetermined tasks (Torrance & Pryor, 1998). Wiggins (1998) is of the view that students need to be given the same training that the assessors receive, so as to be able to judge whether their work is up to standard. Despite lack of understanding of principles of assessment by teachers, they spend more than half of their professional time involved in assessment-related activities (Stiggins, 1997; Boyle & Christie, 2000). This prompted Pellegrino, Chudowsky and Glaser (2001) to declare that instruction on how students learn and how learning can be assessed should be a major component of teacher pre-service and professional development programmes.

Teacher role conflict

Teacher training in assessment is important, but on its own is inadequate since the teacher is required to play a dual role of facilitator and assessor of his/her students (Keightley, 2002; Keightley & Coleman, 2002). Chong (2009) points out that such a situation subjects the teacher to a serious challenge because s/he cannot suppress one when engaged in the other. As a consequence, one of the most frequent concerns about school-based assessment is the issue of teacher bias (Keightley & Coleman, 2002), and it is not surprising to find low variance and a skew towards high marks (Grima & Ventura, 2000).

Some argue that high marks are expected since students are guided by their teachers during the learning process, and are encouraged to improve their performance before they are awarded the final mark for their work (Maxwell, 2004). The portfolio assessment is the closest example in which initial tasks are given low weightage comparatively (Nitko & Brookhart, 2007), or a few best experiments being chosen and scored, as is the case in Malta (Grima & Ventura, 2000). All these tend to affect some teachers' judgements. In some instances, teachers are affected by physical attractiveness of students, by aspects of behaviour or perceptions of ability (Raffan, 2000), and so award or deny marks where they are due.

Lack of confidence in internal assessment

Stiggins (1995) reported resistance towards introduction of teachers' classroom assessment. There is a widely held perception that any examination where external examination does not feature strongly is unreliable and biased (Broadfoot, 1994; Chong, 2009; Keightley & Coleman, 2002; Stiggins, 1997), and even in countries such as Australia, where performance assessment has been in existence since 1970, this perception is still entrenched (Keightley & Coleman, 2002). Of late, there is a paradigm shift towards embracing performance assessment, although many are still equating assessment to external examinations (Raivoce & Pongi, 2000). Lack of confidence by the public is borne from the public's lack of understanding of basic principles of appropriate test interpretation and use (Pellegrino, Chudowsky & Glaser, 2001).

Plagiarism

Plagiarism is one of the challenges in performance assessment (Pongi, 2004), the most common forms being:

1. word-for-word copying of sentences or paragraphs from one or more sources which are the work or data of other persons (including books, articles, working

papers, conference papers, websites or other students' assignments), without clearly identifying their origin through appropriate referencing.

2. closely paraphrasing sentences or paragraphs from one or more sources without appropriate acknowledgment in the form of a reference to the original work or works.

3. submitting work which has been produced by someone else on the student's behalf as if it were the work of the student.

4. producing work in conjunction with other people (other students, a tutor, parents) when it is purported to be work from the student's own independent research. (http://www.griffith.edu.au/).

The list is not exhaustive, as Maxwell (2004) argues that even work that is refined and resubmitted on the basis of teacher feedback may constitute plagiarism, since it is difficult to separate the student input from that of the teacher. In other examination boards, validation is through authenticating the marks on the final form (Grima & Ventura, 1998), and when in doubt the candidates are called for an interview to establish if the work was copied or recycled. However, it is not always possible for teachers to realize that the work presented is not original.

All these problems consequently lead to low validity and reliability of performance assessment, which is discussed in Section 3. 7.

3.6 THE CONDUCT OF PERFORMANCE ASSESSMENT IN BOTSWANA

The recommendation to incorporate performance assessment in the final grade was made in the First Commission on Education of 1977 (Government of Botswana, 1977), and reiterated by the Second Commission in 1993, resulting in the Revised National Policy on Education (RNPE) of 1994. Following the second recommendation, the Examining Board formed a task force in September 1993, comprising Ministry of Education and the Examination Body Officials, to consolidate needs assessment for basic education in

Botswana. In 1998, the Examining Board engaged a consultant to report on the logistics and modalities of implementing performance assessment.

Both the Task Force and the consultant recommended the introduction of Criterion-Referenced Testing (CRT) with diagnostic capability; development of Continuous Assessment (CA) procedures for all school grades to be used as part of final examinations results; and development of materials and training programmes in CRT and CA during pre-service and in-service teacher training, as well as training Ministry of Education personnel such as Principal Education Officers (PEO) (Nitko, 1998). It can be reported that no further work has been done since, and there is no policy on performance assessment other than subject-specific procedures.

Currently, performance assessment is limited to practical subjects and quality is assured through visiting moderation at the end of the coursework, where the teacher and the moderator reconcile their differences (Radnor & Shaw, 1995). Statistical moderation is applied in Design and Technology, in addition to visiting moderation. External moderation is preceded by internal moderation in case more than one teacher was involved in marking.

The issue of payment for performance assessment as experienced in other parts of the world (Grima & Ventura, 2000) is not exceptional to Botswana. Teachers argue that performance assessment increases their already high workload. The issue is so serious that, in 2008, Teachers' Unions took the government to court, demanding to be paid for conducting performance assessment used for external purposes or it is removed from their mandate. The court ruled in their favour. As of September 2009, Teacher Unions have instructed teachers not to submit performance assessment marks until the Examining Body has agreed to pay (*Mmegi* Newspaper, 26 October, 2009, p. 4). This development negatively affects the public's confidence in teacher assessment for certification.

The weight of performance assessment is very little, ranging from 20% in Science subjects and Agriculture (MoE&SD, 2009), to 50% in the majority of subjects. Only Art and Design is assessed 100% by performance assessment (MoE&SD, 2001). The low weightage compares well to other African countries, such as Kenya, Namibia and Tanzania (Njabili, 1987; Noor, 2008; van der Merwe, 2000). Thobega and Masole (2008) attributed the low contribution by Agriculture performance assessment to its questionable reliability. On the other hand, a study by Rathedi (1987) pointed to the need to increase performance assessment contribution towards the final grade. For example, lecturers (78.7%) and graduates (96.7%) did not embrace the testing mode to an excessive degree at the expense of contextualized learning going on throughout the course of study. Rathedi did not outline how quality would be assured for performance assessment to be valid and reliable.

External monitoring and supervision regarding performance assessment is not sufficient, as a result of confusion as to whose responsibility it is among the four departments[7] of the Ministry of Education in Botswana. Mogapi and Yandila (2001) and Yandila, Komane and Moganane (2003) summarised the problems of teaching and conducting performance assessment in Botswana to be: large class sizes of up to 40 students; large teaching loads; absence of laboratory assistants; lack of exemplary teaching materials; inadequate training to carry out coursework assessment; and insufficiency of teachers' orientation on appropriate teaching methods.

Based on the findings of Sections 3.5 and 3.6, quality assurance processes for performance assessment between Botswana's and International practice are summarised in Table 3.1.

Table 3.1: *Comparison between Botswana and international practice on quality assurance processes for performance assessment*

Characteristic	International practice	Botswana practice
Teacher training	Advanced to the extent that teachers develop their own good quality assessment tasks and procedures,	Not emphasised
Moderation	Result involves an interpretation of the final product of the student's work by a judgement of the standard it demonstrates when compared to a set of grade descriptors. Moderation directed at ensuring quality by using a variety of methods that combine both quality assurance and quality control procedures.	One moderator with an outsider perspective. Moderation is a one-off activity by one person directed at controlling quality at the end of the process.
Accreditation	School are accredited and visited during the year to evaluate the physical and human resources, assessment methods and procedures. Participating schools are required to submit an assessment programme.	Schools are inspected once at the beginning when it applies to offer the subject.
Monitoring & Supervision	Monitoring of adherence to standards is done both internally and externally by visiting schools each year and to assist teachers in the delivery of the learning programmes.	Monitoring internally is not rigorous. External monitoring is not common.
Workload	Small class sizes.	Large class sizes
Development of tasks	Tasks are developed either by teachers after undergoing vigorous training or centrally developed. Tasks are of high quality.	No task frames or centrally developed tasks. Every school develops its own task.
Scoring	It is done by multiple raters.	It is done by one rater
Assessment Instrument	Use of detailed clearly written criteria.	Individual schools or even teachers develop their own
Weight	High	Low

3.7 VALIDITY AND RELIABILITY OF PERFORMANCE ASSESSMENT INTERNATIONALLY

The issue of validity and reliability in performance assessment is topical (Burger & Burger, 1994; Chong, 2009; Cizek, 1991; Kane, 2008; Mehrens, 1992; Messick, 1989). While validity and reliability of standardized norm referenced testing is well established (Stobart, 2008), that of performance assessment is not (Hargreaves, 2007). For performance assessment to provide credible outcomes there should be no compromise on their validity and reliability (Linn et al., 1993; Mehrens, 1992). Given that absolute validity and reliability are almost impossible to achieve even in written examinations (Harlen, 1994), van der Merwe (2000) implores psychometricians to adopt a lenient stance toward accepting lower levels of validity and reliability.

Comparatively, performance assessment has been found to rate highly for all aspects of validity (Linn et al., 1991), but there have often been significant problems with reliability (Broadfoot, 1994). The claim of established validity of performance assessment was countered by Cizek (1991) and Mehrens (1992), who argued that this only applies to face validity, and so pertains only to what the test appears superficially to measure. Since reliability can be more readily evaluated and quantified than validity, reliability is persistently emphasised, even at the expense of validity (Raffan, 2000). However, Woods (1991) and William (1992) adopted a compromise stance of a trade-off between reliability and validity for any national system of examinations.

3.7.1 Validity

Nichols and Williams (2009) purport that the concept of validity has evolved over time from the validity of an instrument (Ary, Jacobs, Razavieh & Sorensen, 2006) to the interpretation, meaning and usefulness of the scores derived from the instrument (Ary et al., 2006; Salvia & Ysseldyke, 1998; Yao, Thomas, Nickens, Downing, Burkett & Lamson, 2008). This evolution was emphasised by Ary, et al (2006) when they wrote: “validity does not travel with the instrument” (p. 243).

Although Lissitz and Samuelsen (2007) are still of the view that validity is the property of the test, independent of any proposed interpretation or use of the results, most textbooks even today talk of the validity of the instrument. This certainly does not apply to validity in qualitative studies, where it is addressed through honesty, depth, richness and scope of data achieved (Mertens, 2010). There are several different kinds of validity, but only a few applicable to this study are outlined, namely internal validity, external validity, content validity, construct validity, criterion-related validity and consequential validity (Cohen, Manion & Morrison, 2000).

Internal validity

It is the intention of any study to maintain a high degree of internal validity, that is, the observed changes in the dependent variable which are due to the effect of the independent variable, and not to some other extraneous or lurking variables (Mertens, 2010). Aiken (1996) asserts that internal validity is akin to reliability (p.65), and in qualitative research it is assured through credibility, dependability, conformability, and authenticity of data (Mertens, 2010). To improve internal validity of the study, threats have to be eliminated (McDavid & Hawthorn, 2006), through strategies suggested by Cohen, Manion and Morrison (2000) which include: triangulation of data collection methods; using participant researchers; using mechanical means to record, store and retrieve data; using peer examination of data, and persistent observation. Furthermore, the authors argued that threats to internal validity in qualitative research are built in, since it is assumed that they will happen.

External validity

External validity refers to the degree to which results can be generalised to the wider population, cases or situations (Aiken, 1996; Cohen, Manion & Morrison, 2000) based on the assumption that the sample is representative of the population (Mertens, 2010). To achieve external validity in quantitative research,, variables have to be controlled, and samples randomized, whilst for qualitative research human behaviour is infinitely complex, irreducible, socially constructed and unique (Cohen, Manion & Morrison, 2000). Yin (2009) suggests that the use of multiple cases can strengthen the external validity of results.

External validity in qualitative research is interpreted as comparability and transferability (Guba & Lincoln, 1989; Lincoln & Guba, 1985), thus data in qualitative research can be translated into different settings and cultures. If clear, detailed and in-depth description of research is made then others can decide the extent to which findings from one piece of research are generalisable to other situations. Lincoln and Guba (1985) and Bogdan and Biklen (1992) caution researchers that it is not their task to provide an index of transferability, but rather to provide thick description (Mertens, 2010) of the settings, people, and situations to which they might be generalised.

Content validity

Whether data is collected using either adopted or adapted instruments or developing one's own, determination of content validity is an important first step (Viswanathan, 2005). In judging content validity, the content domain, which includes both the subject matter and the type of behaviour or task desired from students (Mehrens & Lehman, 1991; Aiken, 1996; Moskal, & Leydens, 2000; McIntire & Miller, 2007), and universe of situations must first be defined, and thorough inspection of the items made. Recently, Kane (2008) redefined content validity to include both judgments about content and some analysis of reliability and scaling issues. It should be emphasized that an instrument may have high content validity for one user and low content validity for another, because they wish to infer to different domains (McIntire & Miller, 2007).

Construct validity

Thorndike and Thorndike-Christ (2010.p.11) define a construct as an abstract. Construct is therefore the most important and most difficult form of validity to establish (McIntire & Miller, 2007; Viswanathan, 2005), hence the construct should be operationalised. Fink (2005) simply defines construct validity as a measure that distinguishes between people who have certain characteristics and those who do not. An instrument is said to have construct validity if the instrument results are in keeping with this expectation (Devitt, Kurrek, Cohen, & Cleave-Hogg, 2001).

There are two strategies for demonstrating construct validity, namely convergent and discriminant (McDavid & Hawthorn, 2006; McIntire & Miller, 2007; Viswanathan, 2005). In convergent validity constructs that should be theoretically related are indeed related, while in discriminant validity constructs that are not supposed to be linked are not correlated (McDavid & Hawthorn, 2006; McIntire & Miller, 2007). As with internal validity, construct validity is also vulnerable to threats, two of which were identified by Messick (1995) in Mehrens (1991) as being major ones: (1) construct underrepresentation, in which the assessment is too narrow and fails to include important dimensions or facets of the construct; and (2) construct irrelevant variance in which the assessment is too broad and contains excess variance because of intrusion of other constructs.

Criterion-related validity

Criterion-related validity is a validation method used to determine whether a test indeed predicts what it claims to predict (McIntire & Miller, 2007; Mehrens & Lehman, 1991). A test has evidence of criterion-related validity when it demonstrates that its scores are systematically related to a relevant criterion. Predictive and concurrent validity are two types of criterion-related validity (Ary et al., 2006; McIntire & Miller, 2007). Predictive method is used to forecast future performance (Fink, 2005) while concurrent related predicts current behaviour (Mertens, 2010) by determining whether scores on a specific test are systematically related to a criterion method collected at the same time as the test (McIntire & Miller, 2007). Concurrent validity finds its most important application when the evaluator has created a new measure that s/he believes is better than the previously validated one (Fink, (2005).

Consequential validity

Consequential validity refers to the social consequences of test interpretation and use (Mertens, 2010). Messick (1995) cautioned that this type of validity should not be viewed in isolation as a separate type of validity, because it is integrally connected with construct validity. The researcher needs to identify evidence of negative and positive, intended and unintended, outcomes of test interpretation and use (Mertens, 2010). To ameliorate particularly the negative unintended outcomes, the test instrument should not miss something relevant or contain something irrelevant that interferes with the affected persons' demonstration of competence (Messick (1995).

The different validity evidence of performance assessment can be enhanced by removing an element of bias from the set tasks. Miller-Jones (1989) argues that "the use of 'functionally equivalent' tasks that are specific to the culture and instructional context of the individual being assessed" (p. 363) could be used to counter the problem of bias. This is because students' past experiences, their interests and the meaning they attach to the task are important factors not to be ignored.

Validity of inferences made of test results could be improved by increasing the number of assessment tasks or using a matrix sampling design, whereby different performance assessment tasks are administered to separate samples of students, during the design of performance assessment programme. Similarly, the extent to which a test's items actually represent the domain or universe to be measured is a very important factor in validating the test use (Moskal, & Leydens, 2000).

The design and development of tasks in collaboration with subject matter experts and stakeholders, particularly practitioners, is an important aspect in validating performance assessment (Burdett & Johnson, 2009). Subject matter experts complement each other in selecting appropriate items and by defining the content domain and universe in terms of both the subject matter and the type of behaviour or task desired from students (Mehrens & Lehman, 1991). Another way of validation could be through authentication of the marks on the final form by schools (Grima & Ventura, 1998). When teachers have not seen the development of the students' work over a period of time, teachers are asked not to authenticate the work. In that case, the candidates are called for an interview to establish whether or not the work was copied or recycled.

There is lack of agreement over how to validate analysis of qualitative research, and thus several contending positions (Lee & fielding, 2009). One example is through the quality of fieldwork, which addresses the adequacy of analysis by reference to factors such as the extent of fieldwork, effort devoted to coding, and the proportion of data accounted for by the most prominent analytic themes (Lee & fielding, 2009). On the other hand, there is validation through ethnographic authority (Hammersley & Atkinson, 1983), which gives credence to the researcher's interpretation since one would have witnessed events unfolding. Others propose that validity can be derived from systematic analytic procedures such as grounded theory (Glaser & Strauss, 1967) or micro-analysis applied to interview data (Agar & Hobbs, 1982). Lately, a postmodern approach to validating qualitative research through analysis which empowers research subjects (Altheide & Johnson, 1994) enjoys wide application. However, there is an increasing acceptance that what counts in establishing validity is the operation of the research community itself (Lee & fielding, 2009).

3.7.2 Reliability

Reliability is concerned with the consistency, stability and dependability of the scores. Popham (2005) contends that, as with validity, the reliability of the instrument is ascertained from the results obtained by administering the instrument. Such an instrument should be free of measurement error and ambiguity (Mertens, 2010), so as to obtain accurate measurements (Fink, 2005). For example, a self-administered questionnaire should be easy to understand, written in simple language at the level of the respondents, and have clear instructions. Reliability should therefore be calculated after every use because it is associated with the interpretation of the scores than with the instrument. Just like validity, there is no fixed reliability coefficient of an instrument (Mertens, 2010).

Reliability can be determined through several approaches. For instance, the *coefficient of stability (test-retest)* involves administering the same test to the same group of respondents on two occasions, with or without time lag (Mertens, 2010). The scores from both administrations are compared to determine the consistency of response. Aiken (1996) cautions that since the conditions of administration are likely to be different over long time intervals from over short ones, the size of test-retest coefficients tend to be larger when retesting takes place after a shorter time than after several months. The test-retest method is appropriate only when test takers are not permanently changed by taking the test, or when the interval between the two administrations is long enough to prevent practice effects (McIntire & Miller, 2007). It is therefore important that whenever reporting the test-retest reliability, the length of time that elapsed between the two administrations should be stated. To circumvent the problem of practice effect, two parallel forms of the same test are given to the same test-takers.

The *Internal consistency* method was devised to overcome problems inherent in the repeated measures (Aiken, 1996; Mertens, 2010). In this type of reliability, a test is given only once to a group of respondents, split into halves before the set of individual test scores on the first half is compared with the set of individual test scores on the second half (McIntire & Miller, 2007). However, for this method to yield an accurate estimate of reliability, McIntire and Miller (2007) propose that the halves be equivalent in length and content. Questions are assigned to each half by random assignment, to balance errors in the score that can result from order effects, difficulty, and content. Since splitting the test shortens test length, hence decreasing reliability, Thorndike and Thorndike-Christ (2010) suggest adjusting the reliability coefficient using the Spearman-Brown formula.

An even better way to measure internal consistency is to compare individual scores on all possible ways of splitting the test into halves using KR-20 (McIntire & Miller, 2007). KR-20 is used to calculate internal consistency for testing whose questions can be scored as either right or wrong, while coefficient alpha is used to calculate internal consistency for questions that have more than two possible responses (Aiken, 1996; Thorndike & Thorndike-Christ, 2010). According to McIntire and Miller (2007), internal consistency is appropriate only for tests that are homogenous, that is those that measure one trait only.

Scorer reliability is concerned with how consistent the judgments of the scorers are (McIntire & Miller, 2007). When scoring requires making judgments, two or more scorers should score the test, using clear instructions for doing so. Scorer reliability can either be inter-rater or intra-rater reliability (Mertens, 2010), the former being concerned with reliability between two independent raters, while the latter compares two data sets scored by the same rater. Score reliability can be expressed as either a reliability coefficient or expressed as a simple percentage of agreement between the two observational data (Mertens, 2010). Fink (2005) suggest that scorer reliability can be enhanced by training data collectors and providing them with guidelines for recording observations, monitoring and discussion of problems encountered by data collectors.

These different forms of determining the reliability of an instrument can be used to reduce the possibilities of lowering the reliability of the study. The discussion that follows outlines some of them.

As discussed above, employing multiple measurements of the same skill mitigates the problem of reliability (Airasian, 2005; Airasian & Russsell, 2008; Crooks, 2004; Linn & Baker, 1996; Maxwell, 2008). Multiple rating could be done by different raters scoring students simultaneously, or the same rater scoring the same student at different times. Multiple raters can improve reliability (Rudner, 1994) because the errors of each observer tend to compensate for the errors of others (Thorndike & Thorndike-Christ, 2010). Multiple rating per se is not the solution to the problem of reliability as evidenced by various studies which have shown that written essays were scored differently by different raters and that even the same rater scored responses differently at different times (Rennert-Ariev, 2005).

The reliability of scoring through multiple rating can be enhanced if criteria are used (Nitko, 2004), and procedures put in place, such as recalibration of raters through refresher practice sessions (Johnson et al., 2009), to avoid 'rater drift' (Becker & Pomplun, 2006). Torrance (1995) and Shavelson et al. (1992) suggest training of observers on using scoring criteria.

Developing tasks of equivalent difficulties (Maxwell, 2004) which can be administered to different students also enhances reliability. Performance on one task provides a relatively weak basis of generalisation to other seemingly similar tasks. The limited degree of across-task generalisability in performance implies that performance needs to be assessed across several tasks. It has been found that increasing the number of tasks is generally more important than increasing the number of raters, (Linn & Baker, 1996).

Because of the need to enhance reliability, some countries use highly standardised tasks and conditions, which have a tendency to reduce the validity of the tasks (James, 1994; Woods, 1991; Lennox, 2000). Others concentrate on aspects which are more readily measurable, such as knowledge and understanding. While this improves the reliability of measurement, it leads to the detrimental neglect of higher-level competencies and attitudes (Christofi, 1988).

The reliability of the results may be influenced by a number of factors, such as respondents' maturity, items that are ambiguous or unclear, and conditions of administration. Reliability of the instrument is improved by developing it in collaboration with stakeholders, thus ensuring that the construct to be measured is succinctly captured. The administration of a test is an important aspect in improving reliability (McIntire & Miller, 2007). Proper administration requires a manual detailing all procedures that all test takers should experience (Johnson et al., 2009).

In a situation in which qualitative data is collected, the researcher as the main qualitative data collection instrument should be sensitive, holistic, adaptable and responsive to changing circumstances, and observe activities silently (Guba & Lincoln, 1981) so as not to influence the outcome, thus improving the dependability of outcomes. Reliability is thus ensured by following an analytic inductive methodology in observation to test emergent propositions (Alder & Alder, 1994). Presentations of observational findings are then written in such a way that the accounts will contain a high degree of internal coherence, plausibility and correspondence to what readers recognise from their own experiences and from other realistic and factual texts (Alder & Alder, 1994).

However, the issue of validity and reliability should be approached with care, as Harlen (1994) points out that these are complementary terms which when one increases, the other becomes more difficult to attain.

3.8 CONCEPTUAL FRAMEWORK OF THE STUDY

The problem of Performance Assessment in senior secondary schools in Botswana, as discussed in Section 1.3, stemmed from lack of policy on continuous assessment resulting in variation in its conceptualisation. Teachers engaged in tasks of non-equivalent demands and non-standardised scoring resulting in the outcomes whose validity and reliability were uncertain. These necessitated an undertaking to understand and explore the characteristics and quality processes essential in the performance assessment of Agriculture Form Four students to ensure valid and reliable examinations in Botswana. Quality assurance processes should be embedded in the system if the outcome is to be reliable. Richard (1993) and Wild and Ramaswamy (2008) consider embedding quality into the processes as a process approach, whereby all the factors that have an impact on the students' achievement are examined. Such factors are found at both system-level and school-level. School-level factors are nested within the system implying that any improvement in the system results in the improvement in the school system.

Figure 3.1 presents factors affecting the validity and reliability of performance assessment. These factors draw from the work of Queensland Studies Authority (1998; 2008; 2009) and Wild & Ramaswamy (2008) in the case of policy formulation; Knostantopolous (2008), Knostantopolous and Chung (2009) for teacher workload; Jones, 2006 and Miller, Sen and Malley, 2007 and Finn et al., (2003) for student-teacher ratio; Stiggins (1997, 2002) and Wiggins (1998) for teacher training; Tindal and Hladyna (2002) and Nitko and Brookhart (2007) for resources provision; Mamary (2007) for school leadership and monitoring and supervision; McMillan (2004) and Popham (2005) for learning autonomy; Deakin Crick, Broadfoot and Claxton (2002) and Harlen (2006) for student motivation; Airasian and Russell (2008) and Nitko (2004) for multiple modes of assessment and multiple rating; and Wiggins (1998) for student readiness for assessment. It has at its centre performance assessment, which is influenced by the both system-level factors and school-level factors. System-level factors include, but are not limited to, assessment policy, monitoring and supervision, student/teacher ratio, teacher training, teacher workload, and provision of resources. On the other hand, school-level factors include school leadership; learning autonomy, student motivation, multiple modes of

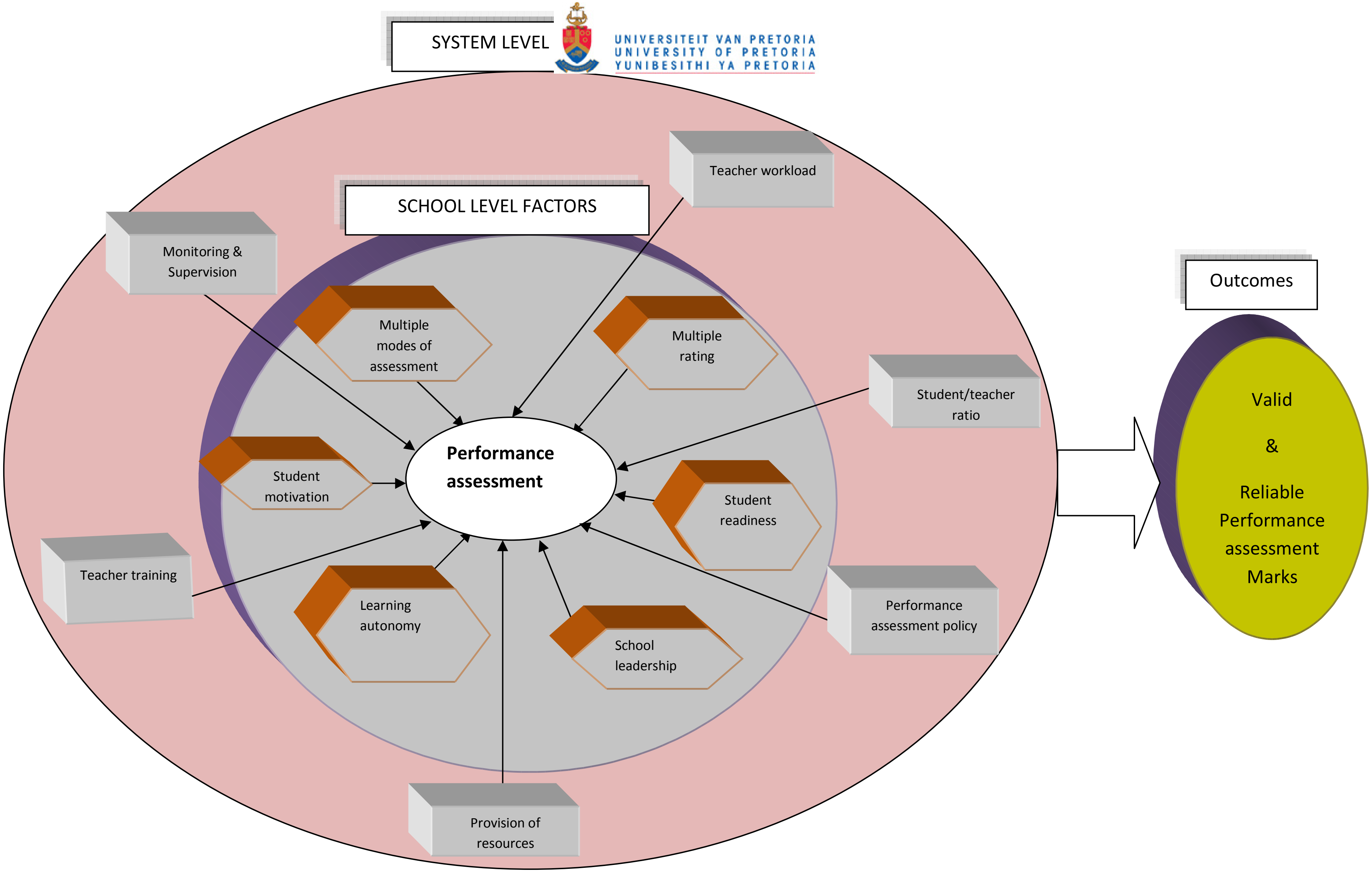

Figure 3.1: Factors affecting the validity and reliability of performance assessment marks

assessment, multiple rating, and student readiness.

3.8.1 System-Level Factors

These are factors that are determined at ministerial level and in most cases a top-down approach is followed. Schools and teachers do not have much say in decision-making. They are required to implement what has been decided for them.

Assessment Policy

Implementing Performance Assessment has to be guided by a policy to produce valid and reliable performance marks for certification (Wild & Ramaswamy, 2008). The policy should outline among other things approval or accreditation based on the quality of teachers to conduct performance assessment, physical and material resources such as tools and livestock. The policy should also spell out objectives that should form performance assessment tasks, who should assess, how many tasks to be done, the roles of the teacher, students, supervisors, and how quality is to be assured (Queensland Studies Authority, 1998; 2008; 2009).

Provision of resources

Implementing performance assessment on large-scale requires massive resources which are costly (Tindal & Haladyna, 2002), just like standardised testing does. Resources such as garden, laboratory, laboratory equipments, tools and equipments, exemplar tasks and assessment materials, and time play a significant role in students' achievement. However, it should be noted that, the presence of equipment and other learning materials do not necessarily imply effective learning and assessment. Performance Assessment by nature requires a lot of time. For example, a portfolio might be developed over a year (Mills, 1996; Johnson et al, 2009). With many students in a class, this might present a mammoth task for the teacher since scoring performance assessment is also a difficult and often time-consuming activity (Nitko & Brookhart, 2007).

Teacher training

Training teachers in assessment methods is vital for successful implementation of performance assessment (Richard, 1993). Teachers are at the forefront of classroom assessment but it was noted by Stiggins (1997) that, "A lot of people involved in education, including teachers do not understand how assessment should be done and why it is done" (p.

2). Training in assessment has never been a prominent part of teacher training (Stiggins, 1997, 2002; Stiggins & Conklin, 1992; Wiggins, 1998) rendering teachers to be unprepared to assess their pupils especially on performance tasks (Kellaghan & Greaney, 2003). For example, Stiggins (2002) noted that "only about fourteen out of fifty states in America require that pre-service teachers take an assessment course. "Only three states require competence in assessment as a requirement for being licensed as a principal, and no state certifies that competence" p. 21.

Despite that, Stiggins (1997) observed that teachers spend most of their time engaged in assessment activities, and yet teacher classroom assessment is weighed proportionally little to the overall mark whenever it is used for summative purposes. However, if teachers are properly trained and given enough support resources they can design and develop sound assessments (Maxwell, 2004). Once teachers have acquired the necessary expertise, they can act professionally and ethically and typically take up the challenges when they are given the responsibility (Maxwell, 2004).

Teachers who lack training in assessment are apt to approach formative assessment in an essentially behaviourist approach. The assessment converges on the assessor's agenda – of trying to find out whether the child knows, understands or can do predetermined things, rather than divergent assessment which emphasises learners understanding (Torrance & Pryor, 1998) and to provide feedback to students on how they have performed in certain objectives and what else they might need to do in order to realize incremental improvement. If we need incremental continuous improvement (Goetsch & Davis, 1997), performance assessment should occupy the centre stage in pre-service and in-service teacher training programmes.

Though in-service training is normally provided, most of the time, initiatives are poorly conceptualized, and insensitive to the concerns of individual participants. Halsall (1998) refers to such designed training programmes as quick fix solutions to the schools' problems. Properly trained teachers in assessment should be able to engage in self-assessment which makes them aware of their own limitations, and those of the techniques they use. If teachers lack competence in some areas of assessment, they should not engage in assessment, no matter how much they are persuaded by the school officials (Salvia & Ysseldyke, 1998). Students too, according to Wiggins (1998) can be given the same training that the assessors receive for them to be able to judge that their work is not up to standards.

Lack of training in assessment causes teachers to unwittingly misinterpret performance assessment and deemphasise or neglect untested material whenever they are engaged in assessing it (Tindal & Haladyna, 2002). Some view school-based assessment (Torrance, 1995) as an extra workload (additional marking, record keeping, and so on) leading to teachers inflating students' marks particularly if they are to be used for summative purposes (Raivoce & Pongi, (2000).

Supervision and monitoring

Monitoring and supervision is extremely important for the success of any project. Supervision must not only be viewed in terms of finding faults on the teacher, but rather as a continuous process aimed at improving teacher performance hence improvement in students learning (Mamary, 2007). Monitoring of school-based performance assessment should be done on daily basis by the senior teacher and routinely by administration. External monitoring is also essential to ensure that teachers do not deviate from standards.

Teacher Workload

Workload normally is positively correlated with class size. The more the students are in a class, the more the work for individualised assistance in a student-centred approach (Knostantopolous, 2008). Workload is probably one of the reasons why teachers ultimately adopt a student-centred approach to instruction and assessment, despite their consciousness of the little impact it has on students' learning of high-order thinking and abstract reasoning. Workload is increased by a lot of recording involved in performance assessment of the student achievement and progress which majority of teachers find it a nightmare (Knostantopolous and Chung, 2009; Torrance & Pryor, 1998). Performance assessment taking place at school level to be included in the certification exerts extra work load on both teachers and students (Fung et al., 1998), because timetabling does not cater for it (Abram, 2008).

Student-teacher ratio

In developed countries, the student/teacher ratio is very small, facilitating individualised instruction and assessment. For example, the G-8 countries' class size ranges from 10 in the Russian Federation to 16 in the United States of America at secondary level (Miller, Sen & Malley, 2007). Jones (2006) reported that in order to increase the connection between materials taught and what students experience in the field setting, the class sizes needs to be

25 or less. If there are just too many students in the classroom, the teacher's assessment focus tends to be on class, or perhaps the small group, rather than the individual student.

Finn et al. (2003) identified reasons as to why and how small classes yield better results. These include: more participation, engagement and identification. There is also more teacher time per student for diagnosing learning problems, working with portfolios, correcting homework, reading with each child and more time for individualised assessment and increased time on task. However, Patchen (2004) and Wiles and Bondi (2000) posit that these can only be effective if teachers change their teaching styles.

3.8.2 School -Level Factors

These are the factors that schools can vary to suit their needs. They include leadership, learning autonomy, student readiness for assessment, multiple modes of assessment, multiple rating, and student motivation.

School leadership

The school head has the duty to manage testing and assessment for the effective running of the school (Mamary, 2007). This includes appropriate conduct of performance assessment which can only be effectively implemented if school management is committed to the responsibility of quality assurance. One of the major functions of management is the formulation, implementation and review of a quality policy (Richard, 1993). If testing and assessment is thoroughly monitored, school policy decisions and instructional leadership/support formulated based on information obtained from quality standard tasks may lead to student achievement hence improved school functional system (Stiggins, 1997).

Supervision is no longer confined to lesson observation (Mamary (2007), but supervisors need to work with teachers on continuous basis and create a school climate in which teachers' self-assessment and co-assessment becomes the culture with the aim to succeed academically. The supervisor's role is to link the purpose and goals of the school to the role of the supervisee and to the improved assessment of the students.

Learning autonomy

The teacher-centred approaches permeating classroom instruction uses direct instruction to whole classes, and appears most applicable: to a well-structured body of knowledge where skills do not follow explicit steps, in introducing and explaining new concepts, in showing how specific pieces of information fit into logical structures and in reviewing and summarising information (McMillan, 2004; Popham, 2005). The need to allow students to actively construct their own knowledge through active participation heralds a paradigm shift. In student-centred didactic, the teacher's role is delegated to explaining basic concepts and skills in facilitating group learning. Recent student-centred learning approaches are based on instructional strategies such as cooperative learning, problem-based learning, discussion, discovery learning, and collaboration. These instructional strategies are feasible in an environment where class sizes are manageable, teachers having skills in assessment, and viable contextual factors promoting effective classroom ecology.

Student motivation

Students should be prepared to continue learning after school. This can only happen if they are motivated to learn (Deakin Crick, et al, 2002). All students want and have the capacity to learn (Greenwood & Gaunt, 1994). The aim of learning is to continually improve their performance and self-esteem, not to measure their failure. Motivation for learning can be fostered in the form of interest, goal orientation, locus of control, self-esteem and self-efficacy, and self-regulation (Harlen, 2006). Motivated students with learning goals have the following characteristics (Dweck, cited in Torrance & Pryor, 1998, p. 85).

- choose challenging tasks regardless of whether they think they have high or low ability relative to other children,
- optimise their chances of success,
- tend to have an incremental theory of intelligence,
- go more directly to generating possible strategies for mastering the task,
- attribute difficulty to unstable factors e.g. insufficient effort, even if they perceive themselves as having low ability,
- persist in their endeavour, and
- remain relatively unaffected by failure in terms of self esteem.

A transparent assessment system where students and the teacher consult each other about assessment; developing rubrics jointly, and applying rubrics to common examples of student work and then discuss the results (Stiggins, 1997; Mergendoller, Markham, Ravitz, & Larmer, 2006), motivates students to achieve.

Multiple modes of assessment

Assessment in educational settings is a multifaceted process, encompassing the way students perform a task in a variety of contexts or settings (Airasian, 2005; Mamary, 2007). As such there are different kinds of achievement to assess which include knowledge, skills, product, reasoning, and dispositional (Stiggins, 1997). Various assessment methods have to be repeatedly employed to reflect those achievements and to allow for all the intended learning outcomes to be appropriately assessed (Maxwell, 2004). Tindall and Marston (1990) identified three sources of information which differ greatly in the type of data and methods employed. These are observations which are non-interactive, interviews which are interpersonal, and testing which focuses on quantifying performance.

Multiple rating

As discussed above, multiple observation of students' performance provides more reliable and accurate information (Airasian, 2005; Airasian & Russsell, 2008). Multiple rating could be done by different raters scoring students simultaneously or the same rater scoring the same student at different times. Raters can score the same piece of work differently and even the same rater can score it differently at different times (Rennert-Ariev, 2005). It is argued that multiple raters can improve reliability of performance assessment just as multiple test items can improve the reliability of standardised tests (Rudner, 1994). The reliability of scoring through multiple rating can be enhanced if the criteria are used (Nitko, 2004), and procedures put in place such as recalibration of raters through refresher practice sessions (Johnson et al, 2009), to avoid rater drift (Becker & Pomplun, 2006).

Student readiness for assessment

The assessment of students is a social act that has social and educational consequences. Students should therefore have the right to the assessment procedures and should have the willingness to complete the assessment. Emphasising the need for student readiness for assessment, Grant Wiggins (1998) had this to say:

> Gone are days when silent examinees sitting in rows, answering uniform questions with orthodox answers in blue books or on answer sheets with No. 2 pencils. Gone are arbitrary calendars that dictate that students must all be examined simultaneously, regardless of readiness (p. 3).

Data generated through assessment is used to make decisions about the students, and the decisions could significantly adversely affect an individual's life opportunities if assessment is improperly made. Salvia and Ysseldyke (1998) assert that those who accept students must accept responsibility for the consequences of their work, and they must make every effort to be certain that their services are used appropriately. Assessment should not be viewed as a means to rank order students to select those who can proceed for further education or employment, but rather to impart social and life skills that they can use outside school.

3.9 CONCLUSION

Performance assessment is being reintroduced by a number of countries as a reform measure in assessment (Khoo & Idrus, 2004; Maughan, 2004), because of its ability to improve learning on the one hand, and its complete evaluation of the student's capabilities on the other hand. Performance assessment includes products and performances such as portfolios, projects, and experiments which when properly implemented results in the acquisition of complex thinking skills, problem-solving skills, and abstract reasoning (ARG, 2006; Shepherd 2000, 2008). These skills give the kind of challenges, diversity and flexibility that make assessment more realistic and educative (Wiggins, 1998) because the thinking processes that students undergo to construct responses is assessed rather than simply auditing learning (Airasian, 2005). Assessing the thinking processes helps students improve their learning (ARG (2002), resulting in deeper and critical thinking.

Performance assessment is engaging, open and uses criteria which describe the expected quality of performance students must be meet (MacMillan, 2000). Feedback is provided to assist students to improve and attain the expected quality of performance. The authenticity of performance assessment allows students to perform in the context of the real-world situations in which the skills are to be applied (Johnson et al., 2009; Nitko & Brookhart, 2007). Whenever performance in real-world is not plausible, simulation is used (McMillan, 2004). Mistakes are made during the trial period and they help students grow and develop over time. Students' growth is influenced by many factors and cannot be assumed to be the same (Neill & Medina, 1992). Though clearly defined criteria are used, an element of subjectivity can creep in when assessment is carried out by one assessor (Thorndike & Thorndike-Christ, 2010). Quite often, performance assessment is used to assess those skills which cannot be assessed by paper-and-pencil which the teacher is best placed to do.

The control of performance assessment by the external bodies varies across countries. Some countries have complete control from determining the curriculum, developing task frames that guide the development of tasks, through developing standardised performance materials, standardising administration, centralise marking to manipulation of the outcomes through statistical moderation (Berry, 2008; Broadfoot, 1994; Lennox, 2000; Singh, 2004). In such countries, performance assessment is used to supplement paper-and-pencil examinations with the consequences of low weightage attached (Kanjee & Sayed, 2008; Maughan, 2004; Singh, 2004; Van der Berg & Shepherd, 2010).

In some countries, performance assessment has completely replaced one-off final examinations (Broadfoot, 1994; Gasemann, 1993; Queensland Studies Authority, 1998). The development of assessment tasks and procedures is the responsibility of the teacher, and the resulting outcome involves an interpretation of the final product of the student's work by a judgement of the standard it demonstrates when compared to a set of grade descriptors (Mercurio, 2008; Maxwell, 2004).

Moderation of the assessment throughout its conduct is the way to ensuring valid and reliable outcome of performance assessment. Subjecting marks to statistical moderation to normalise them when little is known about how the marks were produced does not help in improving the quality of assessment. As indicated above, this results in low weightage being attached to such assessment mode which has the adverse effect of lowering teachers' morale. Since teachers spend more time in this kind of assessment, they expect a corresponding high weight to be given to it. Emphasising training of teachers on performance assessment can result in more valid marks being produced hence increasing the weight of performance assessment to the final mark.

Performance assessment by nature is valid because it represents the actual activities in real life. Properly crafted performance tasks can therefore rate highly in validity when guiding principles are followed during their construction and scoring. Since a score from a single assessment is not reliable, performance scores are generated over a period of time, using different raters. Reliability of performance assessment is low particularly when the stakes are high and teachers tend to collude with both students and parents to inflate students' marks so that they can pas examinations.

Meta-analysis of a number of countries on performance assessment implementation has revealed similarities and differences, particularly with regards to quality assurance procedures. Performance assessment in developed countries emphasises quality assurance procedures, whereas African countries seem to lag behind in emphasising quality assurance aspects due to high level of costs and administrative complexities associated with performance assessment. A number of African countries are still battling with access to education hence quality assurance is given secondary treatment. Thus, conceptualisation of performance assessment followed by entrenching quality into the system is of paramount importance for its success.

Botswana as an African country has to overcome the same problems of performance assessment implementation and entrenching quality assurance processes in the system. Currently emphasis is on quality control carried out once at the end of the year, either by visiting moderators who go out to schools to dictate what teachers should do or applying statistical moderation to the teacher marks which are not necessarily valid. There is no literature on how quality assurance is assured in performance assessment in Agriculture in Botswana schools. Anecdotal evidence suggests that there is quality control at the end of the year in Agriculture performance assessment by moderators. This study will take a step forward to understand and explore the characteristics and quality assurance processes needed in performance assessment of Agriculture Form Four students, and develop quality standard task and assessment materials for use in a quality embedded environment.