

https://dspace.org/

An analysis of the measurement of the progress in learning outcomes at country level: the case of South Africa

Deliwe, Carol Oluyemi Nuga

2017

https://hdl.handle.net/10539/25944

Nuga Deliwe, Carol Oluyemi (2017) An analysis of the measurement of the progress in learning outcomes at country level : the case of South Africa, University of the Witwatersrand, Johannesburg, https://hdl.handle.net/10539/25944

Downloaded from a DSpace repository. https://dspace.org/

AN ANALYSIS OF THE MEASUREMENT OF THE PROGRESS IN LEARNING OUTCOMES AT COUNTRY LEVEL: THE CASE OF SOUTH AFRICA

Carol Oluyemi Nuga Deliwe

A research report submitted to the School of Education in the University of the Witwatersrand, in partial fulfilment of the requirements for the degree of Master of Education.

First submitted: March 2017
Revised submission: October 2017

ABSTRACT

This study contributes to the literature on the development and implementation of sample-based systemic learning assessment programmes which are used to measure the progress in learning outcomes in schooling systems. The justification for focusing on sample-based assessment is for reasons of cost and the need for test-security – conditions which prevail in most developing countries. The study modifies and emphasises the technical aspects of an existing framework, which classifies assessment systems by levels of development. This modified framework and modified rubric arising from the framework are then used to analyse and evaluate the dimensions of enabling context, system alignment and assessment quality of South African learning assessment programmes intended to measure learning progress at country level. Programmes examined include the Annual National Assessment (ANA), the National School Certificate (NSC) and the Southern and Eastern African Consortium for Monitoring Education Quality (SACMEQ) programme.

This study contributes to the body of knowledge on large scale learning assessment at country level. The specific research contribution of the study includes a modified framework and evaluation matrix for analysing educational assessment programmes for measuring learning progress at the country level. The second research contribution is a detailed and updated chronology and profile of these programmes in South Africa. The research and policy implications of the findings of the study include detailed technical specifications for strengthening the measurement of the progress of learning at the country level, drawing on best practice and lessons learned from South Africa's rich and varied participation in learning assessment programmes.

Key words: Standardised Assessment, Large Scale Assessment, Education System Assessment, Educational measurement, Testing, Monitoring Learning Outcomes, Annual National Assessment, South Africa

DECLARATION

I declare that this research report is my own original and unaided work.

It is submitted, in partial fulfilment of the requirements for the degree of Master of Education, to the School of Education in the University of the Witwatersrand, Johannesburg. I declare that I am the sole author of the work and that, except where indicated explicitly, the reproduction and publication of this work by the University of Witwatersrand will not infringe of any third party rights. I declare that I have not previously, in part or in entirety, submitted this report for any degree, examination or qualification at any other University.

_____

Carol Nuga Deliwe

8 March 2017

ACKNOWLEDGEMENTS

I dedicate this thesis to my parents who have always been loving and supportive, and to Taiwo Olusa who has always been a special part of our family. Thanks to my siblings, Joyce, Seye and Sola, for their love, support, humour and our daily on-line discussions. I am particularly grateful for the 'triplets' who steadfastly monitored the progress of my chapters until I completed the study.

Special thanks to Hannah, Lorraine, Motlakapele, Sudeshan, Teresa, Vasu for being my support system; and to Sola, Gugu and Mpumi, my fellow students and personal accountability system. Thanks to my dear friends, aunts, uncles, brothers and sisters who have adopted me – you have given real meaning to the concept of extended family. I would not have finished this without your support.

I would like to express my sincere gratitude to my colleague and friend, Dr Martin Gustafsson, for his guidance, literature and encouragement in this research study. To my professional and policy mentors in one of the most challenging, ambitious and progressive education reform projects in the world, Messrs Mseleku, Soobrayan and Mweli, I am profoundly grateful for your time, advice and commitment to public education.

Finally, I am indebted to my academic supervisor, Professor Brahm Fleisch, for his insightful questions, support and advice which assisted in the completion of this study.

Ese adupe o.

Table of contents

Chapter 1 Introduction and background 7
1.1 Education quality policy developments and the role of assessment 9
1.2 Policy context of educational assessment in South Africa: from 1994 to 2015 13
1.3 Rationale for the present study 16
1.4 Research questions 18
1.5 Chapter outline 18
Chapter 2 Literature review and evidence for the use and components of effective systemic learning assessments, and analytical approach. 20
2.1 Definitions of education, learning, assessment, measurement and evaluation in education 22
2.2 Definitions of formative and summative assessment 25
2.3 Scope, delimitations, and parameters of the study 26
2.4 Examinations and the learning assessment continuum 27
2.5 Use and impact of assessment data on education policy 29
2.6 Test-based accountability, academic performance and assessment-related behaviour 31
2.7 Critique of testing: issues and considerations 37
2.8 Frameworks for analysing effective student assessment systems 40
2.9 Evidence for the technical components, dimensions and standards of effective standardised systemic learning assessments 49
2.9.1 Enabling conditions and context 50
2.9.2 System alignment 53
2.9.3 Assessment quality 56
2.10 Conclusion 74
Chapter 3 Evaluation and analysis of systemic learning assessments in South Africa.. 76
3.1 Analysis of systemic learning assessments in South Africa: ratings and evaluation method 76
3.2 Analysis of systemic learning assessments in South Africa: discussion 98
Chapter 4 Review of past systemic learning assessment programmes in South Africa 106
4.1 Chronology of systemic learning assessments in South Africa, 2016 107
4.2 Description of systemic learning assessments in South Africa, 2016 114
4.2.1 Enabling context 114
4.2.2 System alignment 121
4.2.3 Assessment quality in the original framework by Clarke (2012b) 126
4.2.4 Assessment quality (modifications) 133
Chapter 5 Conclusion: Analysis of systemic learning assessments in South Africa 140
5.1 Policy recommendations 142
5.2 Opportunities for further research 149
5.3 Limitations 151
5.4 Conclusion 153
Chapter 6 References 155
Appendix. Letter of authorisation from Department of Basic Education for this research study **..... 175**

List of figures

Figure 1: Components of an education assessment system 23
Figure 2: Examinations and learning assessment continuum 29
Box 1: Education System Evaluation: Informing Policies for System Improvement....41
Box 2: PIRLS Comparisons possible across years and groups 117

List of tables

Table 1. Analytical themes used in developing an overview of national education systems (reproduced from Ferrer, 2006) 43
Table 2: The development stages of a student assessment system 45
Table 3 : Development levels for different types of assessments 46
Table 4: Development levels for National Large Scale Assessments (NLSA) 47
Table 5: Time taken to release assessment reports on South Africa's participation in various learning assessments 66
Table 6: Evidence for Clarke's original framework and modifications to the SABER rubric (2014) for analysing programmes intended to measure learning progress at the country level. 73
Table 7: Dimensions and sub-dimensions of systemic learning assessment programmes with modifications arising out of the findings of this study, modified from evaluations rubric developed by SABER and Clarke 77
Table 8: Evaluation matrix with rating rubric and criteria for evaluating the level of development of systemic assessment programmes in South Africa 79
Table 9: Summary matrix with ratings for the level of development of systemic assessment programmes in South Africa. 97
Table 10: Chronology of systemic learning assessments in South Africa, 1994 to 2016 107

Chapter 1 Introduction and background

Given the social, economic and political benefits of education (Burgess, Wilson & Worth, 2013), the quality of learning outcomes has dominated the growth and development narrative in recent years. The contribution of quality basic education to the wealth of nations, society and development is well acknowledged in the literature of various disciplines. The educational, economic and development literature acknowledges the link between the quality of education and the development of nations and individuals, and the elimination of poverty, inequality and social exclusion (Burgess et al, 2013; UNESCO, 2015). The compelling evidence for the importance of basic education through schools provides the rationale for strengthening efforts to better measure the progress and quality of education at the country level.

The evidence for the benefits of education ranges from the human capital theories arising from the work of Mincer and Schultz in the 1950s and 1960s to the more recent capability approaches to education and training advocated by Sen (2003) which confirm the social choices and freedoms that education allows among individuals and within countries. Education has a positive effect on individual income and social mobility, particularly when measured by what learners can do rather than by their enrolment and participation rates[1].

Despite this evidence, and regardless of how learning is measured, the educational endowment in many developing countries still leads to low and inequitable levels of

[1] Research by Hanushek and Kimko (2000), using international assessment data from a wide range of developed and developing countries, shows that the quality of learning is a more influential determinant of individual income and broader economic growth than participation rates. They also found that one country-level standard deviation, equivalent to 47 test score points in PISA 2000 mathematics, is associated with a 1 percentage point increase in annual economic growth rates. In addition, they conclude that the strength of this relationship dwarfs that of the association between participation rates and growth rates. In South Africa, Van der Berg (2008) identified school management and institutional dysfunction as determinants of performance, contributing to the inability of schools serving learners from poor communities to convert resources into learning. Using SACMEQ data, Van der Berg (2008) concluded that resources matter conditionally on the required ability to carry through the conversion, and, significantly, that SES influences outcomes more in SA than in any of the country's regional neighbours especially in less well-resourced schools. Van der Berg (2008) concludes that poor learners suffer conditions that condemn them to poor social mobility through exposure to poor quality education, life and work opportunities.

learning outcomes. This has severe consequences for social and economic development in these countries. South Africa, also, suffers from historically low and inequitable academic achievement at school level which is evident even in the earliest grades (Fleisch, 2008). This is despite the progressive funding regime, prioritisation of public investment in social services and the expansion in school participation observed over the past two decades in the country. Thus, measuring progress in learning is an essential part of monitoring development at country-level in education systems. In this study, the measurement tools for doing this are referred to as systemic learning assessments[2] since this is the terminology used in South Africa although they are also referred to as large scale assessments in the literature (Lockheed, 2008; Greaney and Kellaghan, 2008). Given the importance of schools as sites of learning, monitoring learning process can help countries to deepen and accelerate the acquisition of knowledge, skills and values in their education systems.

The objective measurement of the achievement of learning outcomes is typically achieved through the administration of learning assessment tests, which provide valuable information on academic achievement. The results of these assessment tests are generally taken as objective measures of acquired knowledge, skills and competencies mastered by learners[3] although there is substantial contestation about the unintended consequences of assessment testing and the degree to which assessment tests can be used to measure all the necessary dimensions of learning considered to be important. Test design and test results, which carry significant social, political and economic weight in the development policy narrative and they are frequently the subject of intense public discussion, policy debate and professional contestation in schooling systems across the globe (Brookhart, 2013).

High performing education systems manage to achieve alignment, accountability and stability in what Slavin (2005) refers to as an orchestra of teaching and learning in

[2] According to Ravela et al. (2009) cited in Clarke (2012b), an assessment system is a group of policies, structures, practices and tools for generating and using information on student learning.

[3] The term 'learner' is defined in the South African Schools Act (1996) as any person receiving education in terms of the Act. 'Learner' is synonymous with 'student' or 'pupil' in other education systems.

the complex enterprise of education. This study focuses on tracking the progress of learning in schools at the country level. Specifically, it examines the international experiences of technical implementation of monitoring systems, analyses the nature of assessment system components within programmes at country-level and proposes how tracking the measurement of progress in learning outcomes may be improved in South Africa in support of systemic education quality improvement. Systemic learning assessment refers in this study to learning assessment in schools for the purposes of monitoring the progress in education system performance at country-level.

The intention of the study is to contribute to the policy debates which inform investment in assessment in the schooling system. The focus, for reasons of cost, is on sample-based systemic learning assessments rather than on universal assessments although, as will be seen in Chapter 4 and 5, many of the technical recommendations of the study relate to other types of educational assessment.

This chapter presents the background, context and rationale for the study and reflects on the national and global developments in, and context of, education system reform. The role of assessment testing in measuring progress in education is introduced and further elaborated in the subsequent chapter. The gaps in the current literature on the technical implementation of learning assessments provide the motivation, and the main research questions are discussed. The research questions provide the basis for the main work of the study as they are used to guide the examination of best practice in systemic learning assessment internationally; to evaluate South Africa's assessment efforts at the country level; and to develop policy and research recommendations for improving the measurement of progress in learning outcomes in the country. The chapter then concludes with an outline of the dissertation, a summary of the areas covered in the literature review, and a discussion of the analysis and findings of the study.

1.1 Education quality policy developments and the role of assessment

Over the past two decades, the global development agenda, aided by the Education for All (EFA) movement, has evolved from the 1990s preoccupation with school

participation to a focus on learning outcomes. This is demonstrated in the development of *Sustainable Development Goal 4* on Quality Education, part of the *2030 Agenda for Sustainable Development*. The development of these Sustainable Development Goals emanated from a global assessment of education progress. This confirmed that the expansion in global access to schooling over the past three decades was not accompanied by an improvement in educational outcomes in developing countries (UNESCO, 2015a).

Various researchers have explored the concept of test score use in school accountability. Levitt, Janta, & Wegrich (2008) in their literature review commissioned by the General Teaching Council for England (GTC) informed the development of proposals for a new accountability framework for teachers in England. They identify the ethical origins of accountability involves the appropriate and proper behaviour in relation to responsibilities between agents responsible for actions and principals holding them responsible, where the actions are effected between individual, institutional and social actors as defined by Bovens (2005). Figlio and Ladd (2007) refers to school accountability systems which use test scores to hold schools accountable although Koretz and Hamilton (2003) describes challenges with this approach, while noting the increasing tendency to use such scores to hold those within schools accountable with extreme consequences. Levitt et al (2008) identify the use of the external accountability model (also referred to as bureaucratic or hierarchical accountability in the literature) since internal accountability is difficult to quantify. In using external accountability model to understand the progress in learning in an education system, Rosenkvist (2010) defines schools as an instrument for education policy on the national, regional and local level, and shows how they are frequently compelled to provide information to policy makers and the public about value for money, compliance with standards and regulation and quality of the services provided. Schools and teachers may generally be held accountable for the quality of the education they provide, using the external accountability model.

The rise in evidence-based and accountability-focused reforms over the same period explains the increasing interest in, and appetite for, measuring the quality of learning

outcomes.[4] These include the standards-based accountability reform movement which originated in the US in the 1990s, and the evidence-based policy reform approaches adopted by different sectors of the United Kingdom (UK) government around the same time (Sutcliffe and Court (2006) in Best, Knight, Lietz, Lockwood, Nugroho & Tobin (2013)).

Many different terms are used in the literature on educational performance, achievement and outcomes, so it is necessary to clarify the terms used in this study. An educational assessment system in a country includes four types of assessment programme. Clarke (2012b) proposes that these types are examinations (typically summative at the point of exit from schooling); classroom assessment (provided in support of instructional improvement for formation of knowledge, skills and values); international large-scale assessments (typically sample-based and administered to more than one country for the purposes of benchmarking or comparing learning achievement and educational performance); and national large-scale assessments (these assess national education provision and can be used to measure progress in learning against national learning standards). It is this last form of assessment with which this study is concerned and which is referred to for simplicity as systemic learning assessment. In this study, the term "*schooling system"*, which is more frequently used in the literature, is synonymous with "*basic education system".* For brevity, the term *"education system"* is used to refer to the basic education system. The term "*systemic assessment",* also used for reasons of brevity, refers to "*systemic learning assessment*" carried out at the country level in a basic education or schooling system. The terms "*systemic learning assessment*", "*national learning assessment*" and "*national large-scale assessment*" are used interchangeably in the literature to refer to systemic assessment programmes to monitor the progress of learning outcomes at the country level. The term *"learner"* and *"student"* are synonymous and are used interchangeably throughout this report since the former is

[4] According to Loveless, Costrell & Cuban (2005), standards-based education reform comprises three main activities: defining curriculum standards or what learners should know in terms of their cognitive skills and competencies; measuring learning achievement against these standards; and ensuring that the results are consequential in order to drive improvement through quality-seeking behaviours and activities.

used more frequently in South Africa than the latter in relation to school-going children.

Despite the considerable literature on assessment, there is relatively little on the technical requirements for implementing systemic learning assessments in developing countries. This study contributes to the literature by isolating the technical components of credible systemic learning assessments and providing evidence for their use and application in measuring the progress in learning outcomes at the country level. The study expands and develops an existing framework for classifying student assessments of different types (Clarke, 2012a; 2012b) and, using the modified framework, provides an analysis of systemic learning assessments in South Africa. The resulting analysis and evaluation of the strengths and weaknesses of South Africa's systemic assessment system is used as the basis for recommendations for improvement of such measurements at the country level. The particular focus on sample-based learning assessments in this study is for reasons of cost and security of administration, as sample-based assessments are generally less costly than universally administered assessments, although many recommendations apply to other types of educational assessment.

This study will benefit those interested in and involved with planning, resourcing and investing in assessment systems for monitoring basic education systems. The study's recommendations draw on country experiences of a variety of national, regional and international assessments including the *International Association for the Evaluation of Educational Achievement* (IEA) *Trends in Mathematics and Science Study* (TIMSS)[5], the *Progress in International Reading Literacy Study* (PIRLS)[6], the Southern and Eastern African Consortium for Monitoring Education Quality (SACMEQ) regional assessment programme and the *Programme for International Student Assessment* (PISA), first implemented in 2000 by the Organization for Economic Cooperation and Development (OECD).

[5] Started in 1994.
[6] Started in 2001.

1.2 Policy context of educational assessment in South Africa: An outline of policy from 1994 to 2015

The findings of this study confirm the conclusions of Ravela (2005) whose research indicated that a country's educational assessment systems are influenced as much by its history and by current political, economic and structural arrangements as by the relationships within and between different players, components, levels and institutions in the system. The framework used in the study aims to accommodate some of these dynamics. The study thus contributed to explicitly determining some of the technical considerations and dimensions that should be included in the measurement of the progress of learning at the country level.

To contextualise the study, this section provides insights into curriculum and assessment reforms in the education system of post-Apartheid South Africa arising from documents and data in the education system (see Appendix). These have focused primarily on access, equity, efficiency and quality of education in a concurrent system of government where national government is responsible for policy, standards-setting and oversight of policy implementation and provinces for strategy, implementation and direct provisioning of basic education.

The document *Policy on Education and Training* by the African National Congress (ANC) emphasised equity and redress (ANC, 1994). According to Ndhlovu, Sishi, & Nuga Deliwe (2006), between 1994 and 1999 education and broader reform programmes focused on restorative justice, resources and financial equity. The strong focus on the distribution of resources in schools was based on the findings of the specially commissioned 1996 Schools Register of Needs (HSRC, 1996). The Norms and Standards for School Funding policy (DoE, 1998b) set in place a pro-poor funding framework for schools. The aim of the policy on teacher rationalisation and redeployment of educators was to correct the inequitable distribution of educators between schools, although the voluntary retrenchment package offered to teachers as part of the policy resulted in the unintended loss of many skilled and experienced teachers (Gordon, 2009). Education reform efforts also prioritised the creation of a single national education system and the first *White Paper on Education and Training* (DoE, 1996) allowed for the alignment and integrity of provincial and national

administrations, with districts emerging as the administrative level of interaction with schools. The *National Education Policy Act (1996)*, focused on structures to enhance intergovernmental alignment and the *South African Schools Act* (1996) allowed for the establishment of democratic school management and governance processes and for the coexistence of independent and public schools in the education system.

According to Muller (2004), the only systemic assessment instrument in the pre-Apartheid system was the matriculation examination, which was inconsistent in quality, with little emphasis on school based formative assessment, and which was administered in different forms in up to 15 different education departments using a fragmented curriculum framework (Lubisi & Murphy, 2002). After 1994, continuous assessment was declared to be policy in the *Assessment Policy for General Education and Training* (DoE, 1998a) in addition to sample-based assessment of learning. Within the policy, Systemic Evaluation (SE) was developed. It was the pre-cursor of what emerged as the Annual National Assessment[7]. The outcomes based curriculum approach was borrowed from the United Kingdom (UK) and other industrialised countries, and was articulated in a new curriculum policy called Curriculum 2005 (C2005). C2005 was found to rely too much on highly individualised assessment which was difficult to implement (DBE, 2009) and created a considerable administrative burden in the teaching corps. C2005 was later reviewed, renamed the *Revised National Curriculum Statements (*RNCS) and finalised in 2002 with more specification around the nature, frequency and expectations of assessments, and the subsequent introduction of Common Task Assessment policies (DoE 2002).

Between 1999 and 2004, the focus of education reform was on consolidating assessment policy and curriculum and school enrichment in South Africa (DoE, 1998a, 2002). After Socio-economic status was indicated as the main contributor to learning achievement (Van der Berg and Shepherd, 2010), school management and leadership were identified as major contributors to learner performance in South Africa (Crouch and Mabogoane, 1998), supported the findings of later research on

[7] The 2011 ANA used instruments used in earlier systemic evaluation survey cycles.

school effectiveness and dysfunction by Van der Berg (2008). Various interventions and programmes were introduced to bolster leadership including the establishment of training initiatives for school governance and leadership development[8], the development of the Advanced Certificate in Education (ACE) and the launch of the Whole School Evaluation Policy which aimed to improve school functionality, management and governance processes (DoE, 2001; Reeves, 2005). The reforms in governance, curriculum, administration and equitable resource allocation were part of the Tirisano implementation programme (DoE, 2004), along with enrichment initiatives in history, values and social cohesion in schools.

From 2004 to 2009, the focus shifted from improving access and participation to deepening institutional effectiveness. The *Ministerial Report on the Costs, Resourcing and Financing of Education* (DoE, 2003a) emphasised the need to compensate for the burden of poverty on education and recommended the introduction of no-fee schooling to relieve the burden on households of school-related expenses such as uniforms and fees. The report proposed the development of performance indicators taking account of socioeconomic status and school functionality. Umalusi, the Quality Assurance body for General and Further Education and Training, established in 2002 to oversee quality assurance in the provision of school qualifications and certification, was instrumental in the development of a single Grade 12 National Senior Certificate (NSC) examination in 2008. Between 2010 and 2014, the Curriculum and Assessment Policy Statements (CAPS) were developed and implemented to address curriculum shortcomings identified in three successive curriculum reviews. Previous reviews were concerned about the weak specification of the curriculum and ineffective assessment practice in schools, particularly in those serving learners from poor households[9], resulting in an emphasis in recent years on school- and classroom-based assessment quality in recent education sector plans (DBE, 2015a).

[8] The Matthew Goniwe School of Leadership and Governance was established in 2003.

[9] Interview with Dr R Poliah, Chief Director: National Examinations and Assessment, DBE, September 2014.

The fourth and fifth post-Apartheid governments (2009 to 2014 and 2014 to 2019) have focused on outputs (indicated by the number of examination passes, for example) and outcomes (actual knowledge, skills and competencies acquired). In 2008, Government introduced the Foundations for Learning Campaign, which focussed on the Foundation and Intermediate Phases and included clearer specifications of the materials, time and different learning activities in a week. The campaign involved a number of teacher training initiatives and trial runs of a new national assessment were run in 2008 and 2009 to expose teachers to innovative assessment and other materials (DBE, 2011a).

Following the general election in April 2009, a comprehensive set of activities and targets in the sector were included in the agreements signed by the President with his Ministers in different sectors. The priority government outcome of improved quality of basic education enabled the prioritisation of monitoring of progress against a set of quality-related indicators covering early childhood development, learning and teaching outcomes, materials provisioning, resourcing and management, assessment, learner well-being and school safety. Enabling the strengthened measurement of learning outcomes through effective functional assessment systems is currently articulated at national and sectoral level in South Africa through Chapter 9 of the National Development Plan (NPC, 2012) and the Basic Education Action Plan (DBE, 2010a; DBE, 2015a).

1.3 Rationale for the present study

Research by Ferrer (2006), Braun and Kanjee (2006), Lockheed (2008), Greaney & Kellaghan (2008), and Conley (2015) all confirm that assessment system reform must be part of broader educational reform, in order to ensure quality education for all. Although there are few frameworks to analyse assessment systems for measuring learning progress at the country level, the main ones in the literature are informed by the experiences of countries (Ferrer, 2006; Greaney and Kellaghan, 2008; Clarke, 2012a; OECD, 2013). South Africa's National Development Plan (NDP) identifies the Annual National Assessment (ANA) as a tool for monitoring trends in learning, targeting support, providing better feedback to parents and rewarding school

performance referring to the sample-based systemic assessment function (V-ANA) and the separate universal assessment (U-ANA) which can be tailored to diagnosing learning weaknesses in classrooms (NPC, 2012). The sector plan for basic education echoes these views (DBE, 2015a) and motivates for better measurement of the progress in learning outcome at the country or system level. It must be stated up front that the systemic learning assessments which this study focuses on provide limited opportunities for feedback that would be helpful at classroom and school level, since its benefits are for system-level or country-level monitoring. Likewise, although school-based and classroom-based assessments are associated with feedback on learning progress and opportunities for instructional improvement (Heritage, 2008), these assessments will not be dealt with in detail in this study. Chapter 2 describes some features of such assessment systems, tools and programmes in an affective and comprehensive assessment system for monitoring country level learning progress drawing on the literature and best practice as evidence.

Clarke's framework for analysing the level of development of different types of assessment in an effective education system (Clarke, 2012a, 2012b) provides the basis for the evaluation of South Africa's systemic assessment programmes. The detailed specifications of the technical dimensions of assessment quality in Clarke's framework are contained in an evaluation matrix (Systems Approach for Better Education Results Student Assessment, 2014) which in turn, was modified as a result of this study. The matrix was modified in respect of its assessment quality and system alignment dimensions, using the findings of the literature review. The modified evaluation matrix (SABER, 2014) was then used to analyse weaknesses and strengths of systemic learning assessment programmes in South Africa from 1994 to 2016.

1.4 Research questions

This study addresses the following questions in relation to sample-based systemic learning assessments:

- What is known about the origins, use and utility of country-level learning assessment practices internationally? As noted in the previous section, the term used is systemic learning assessments but globally, these are referred to as national or country-level large scale learning assessments. Some literature refers to National Large Scale Assessments (NLSA).

- Using a modification of an existing analytical framework (Clarke, 2012b) as a point of reference, what are the strengths and weaknesses of the systemic assessment programmes for measuring learning over time in South Africa?

- What research conclusions and policy insights will strengthen the measurement of learning in order to better track progress in learning outcomes in the basic education system in South Africa?

1.5 Chapter outline

The current chapter provides the introduction, background and context, rationale and main research questions. Chapter 2 presents the scope, main definitions, terms, concepts and evidence from the literature on best practice in measuring learning progress at country level and informs the approach, analytical framework and modifications to the framework and evaluation matrix used in the study. The chapter includes a discussion of the relationship between assessment, testing and academic achievement; insights into the consequences of examination and assessment performance; information about the impact of assessment on education policy; and a discussion of the critique of, and likely behavioural and systemic responses to assessment- and test-based accountability in schooling systems. The chapter concludes with a description of various frameworks for analysing assessment

systems in education at the country level (Ferrer, 2006; Braun and Kanjee, 2006; OECD, 2013; Clarke, 2012a; 2012b) with reflections on the need to contextualise the contestation around educational assessment testing in developing countries such as South Africa.

Chapter 3 presents the analysis of systemic learning assessments in South Africa based on Clarke's modified framework and the modified SABER evaluation matrix in relation to enabling context, system alignment, and technical quality of the assessment. The modified evaluation matrix (SABER, 2014) uses a ratings scale to categorise the dimensions and sub-dimensions of the programmes from latent to emerging, established and, advanced levels of development. The analysis makes it possible to identify the strengths and weaknesses of past and current systemic learning assessment programmes, and provides a basis for specifying what countries should consider when they wish to invest in implementing effective programmes to effectively measure the progress of learning in a country's education system. The chapter concludes with a discussion of the findings of the analysis.

Chapter 4 presents a chronological and narrative profile of the various systemic learning assessment programmes in post-Apartheid South Africa in narrative form, drawing on the literature, findings and analysis in the preceding chapters.

Chapter 5 concludes with a summary of the aims and objectives of the study and provides evidence showing that the research questions have been answered. The chapter also describes methodological and other limitations of the study, and makes recommendations for policy and further research, with the suggestion that an existing sample-based assessment (namely the Annual National Assessment) should be refined using lessons learned in the systemic assessment programmes implemented since 1994, and capitalising on the strengths of these in respect of the curriculum alignment of any country-level assessment test for measuring progress in learning outcomes.

Chapter 2 Literature review and evidence for the use and components of effective systemic learning assessments, and analytical approach.

This chapter presents the scope, main definitions, terms, concepts and evidence from the literature for the approach and analytical framework used in the study. The chapter also provides evidence, from the literature, on the best practice in implementing effective programmes for measuring learning progress at country level. Evidence for the modification of both the analytical framework and the evaluation matrix used in this study are also presented.

The literature review covers the main studies, grey literature[10], technical documentation and literature on systemic learning assessment from developing countries. A letter of authorisation permitting access to documentation from the Department of Basic Education is attached in the appendix. This chapter provides an overview of the notions of education quality and the role of educational assessment in education reform in the literature. It examines the use and abuse of assessment test-based accountability[11] and reflects on the behavioural implications of testing in schooling systems[12] and different frameworks used to analyse programmes for measuring learning progress at the country level. To provide balance, the chapter concludes with a summary of the issues of contestation in the education assessment and testing literature.[13] This is a prelude to the analysis and review of systemic learning assessments in the subsequent chapters.

Much of the literature on educational assessment is informed by research originating in the US into psychological and educational assessment in the years just after World War I. With US school enrolment outstripping population growth between 1890 and 1918 (Linn, 2001), tests using scientific and psychometric approaches were used to

10 Grey literature refers to unpublished research including government documents, reports, policy briefs, and other briefing documents.

11 Test-based accountability seeks to use standardised assessment test results to influence behaviour in support of improvement (Hamilton, Stecher and Klein, 2002).

12 The consequences of performance in assessment testing may be high and attract sanctions or rewards in which case these are referred to as high stakes tests. .

13 Braun and Kanjee (2006) define systemic valid assessment tests as those which contribute to curricular and instructional improvements by design through enabling enhanced performance in relation to the test constructs being measured.

sort ever-increasing numbers of students by ability for further studies, university enrolment and work opportunities (Conley, 2015). Subsequently, learning assessment tests developed in the psychological disciplines began to be applied in many facets of education and employment including the ranking of schools. Technological advances in processing and analysing results, experimenting with test formats and more efficient standardisation techniques revolutionised the scale of testing and made possible the establishment of the comparative international assessment studies administered under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) in the 1950s. These were the precursor to TIMSS, PIRLS and the National Assessment of Education Progress (NAEP), a systemic learning assessment carried out every two years in the United States.[14]

The literature on education assessment is dominated by experiences in the United States, the European Union and Latin America which account for most of the documentation and literature on system-wide education assessment reform initiatives in the last two decades, though the US dominates. In his assessment of 100 years of studies, polls and surveys on the effect of testing on student achievement, Phelps (2012) notes that 65% of qualitative studies, 89% of quantitative studies and 95% of polls and surveys covering well over 7 million individual responses are based on research subjects and participants in the US.

[14] The precursor to TIMSS is the NAEP. In this study, the NAEP is used as an example of an advanced sample-based systemic learning assessment programme. NAEP is well respected by assessment professionals and well publicised in the media.[14] This is due to the rigorous technical methods and features of its administration which are well documented and have been adapted for use in well-known international assessments such as PISA, PIRLS and TIMSS. NAEP was first implemented in 1969 as a project of the United States Department of Education. It is administered by the National Centre for Education Statistics (NCES) within the Institute of Education Sciences (Bandeira de Mello, Bohrnstedt, Blankenship, & Sherman, 2015). The NCES is responsible for developing test questions, administering the assessment, scoring student responses, conducting analyses of the data and reporting the results of NAEP. The technical features of NAEP include high levels of standardisation of content, governance procedures for standardisation, verification, validation, documentation and analysis.[14] Planning for the implementation of NAEP starts five years before the next assessment and process are explained and communicated clearly to different role-players about the nature, utility and limitations of NAEP results for different levels in the education system. State and federal responsibilities, plans and activities are articulated in assessment frameworks, documentation and communications to different stakeholder groupings in print and on-line materials (Bandeira de Mello et al 2015).

The main contestation in the literature on educational assessment centres on the costs, type, frequency, resourcing, instructional adequacy and alignment of educational assessment tests. Some view tests as mechanisms for entrenching unfair and unproductive educational practices such as cheating, especially when performance is closely tied to rewards and sanctions. Although concerns about the negative behaviours and effects associated with testing are acknowledged, the framework used in this study assumes that educational assessment can be used to track learning and improve system level monitoring and country-level accountability, Tests results represent learning acquired and they are therefore viewed as an important indicator of the quality of schooling (UNESCO, 2015).

2.1 Definitions of education, learning, assessment, measurement and evaluation in education

Education professionals distinguish between assessment, evaluation and testing as related concepts in the field of education. This section gives the definitions used in this study. Formal, informal and non-formal learning is recognised in the literature (Labelle, 1982; Schiefelbein and McGinn, 2008) as being acquired through schools, interactions in broader society and in households respectively. Huitt (2011) defines learning as a permanent change in behaviour in response to teaching or instruction, and instruction as purposeful and direct management of the learning process traditionally carried out in schools – the main site of learning outside the home according to Schiefelbein and McGinn (2008). This conceptualisation of schools is in line with the human capital approach to education in which the school is the primary site of education 'production' (Boissiere, 2004). Using this approach, learning outcomes such as acquired learning, skills, values and attitudes are influenced by schooling inputs (infrastructure, teachers, materials, curriculum and other educational resources), learner characteristics (background and ability to learn) and school characteristics (including organisation and functionality of curriculum-related processes including teaching and learning, management, governance and accountability mechanisms).

Clarke (2012b) defines student assessment system as a group of policies, structures, practices and tools for generating and using information on student learning to meet decision-making and policy-support needs. To illustrate the utility of the measurement of learning across the entire education system, this study conceptualises an education assessment system as having nested sub-systems which include student assessment at classroom level, teacher assessment, school assessment and systemic assessment, with the latter comprising national and international large scale assessment efforts and examinations. Figure 1 below shows the centrality of learning outcomes measurement in such an education assessment system.

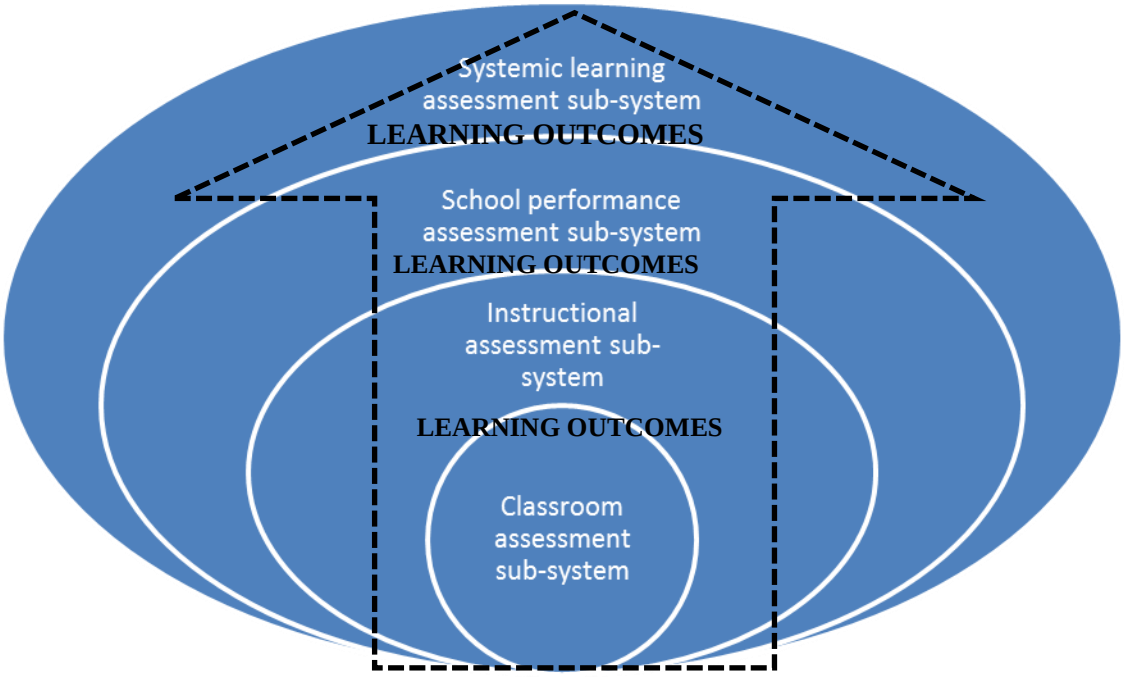


Figure 1: Components of an education assessment system

Source: Author.

Educational assessment involves describing, collecting, recording, scoring and interpreting information about what has been learned (Huitt, Humel & Kaeck, 2001). In this case, the information is about learning achievement in the form of test scores. Measurement refers to the processes and principles against which the attributes of learning are determined. In behavioural terms, to measure is to apply a standard scale or measuring device to an object, series of objects, events or conditions, according to practices accepted by those who are skilled in the use of the device or

scale. Measurements used in educational evaluations include raw scores, percentile ranks, derived scores and standardised scores in assessment tests (Overton, 2011).

Research uses data in order to describe, predict or control the phenomena being researched to generate a better understanding of it. Evaluation involves comparison with a standard in order to judge the quality of the construct being measured (Huitt et al, 2001; Huitt, 2003). Educational evaluation is the judgement about the extent to which curriculum standards have been met through learning programmes. Evaluations typically involve assessment activities that result in information collected about the features, characteristics and outcomes of learning programmes in order to make judgements about the programme or participants in the programme. Research studies may be descriptive, correlational or experimental and more rigorous forms of evaluation such as Randomised Control Trials (RCTs) have been introduced to improve the validity of judgements of the effectiveness and value of large-scale social programmes. However, they require a scientifically determined counterfactual and tend to be costly. However, supporters of scientifically rigorous evaluation methodologies argue that investing in ineffective programmes is a waste of scarce public resources (Mohohlwane, 2016).

Assessment may be carried out to inform parents about their child's performance, inform teachers about instructional and learning weaknesses that need to be addressed, assist in determining student destinations and help policy makers to determine progress in the education system. In classrooms, assessment is considered part of instruction (Black & Wiliam, 1998). At systemic level, it is a research activity intended to establish the levels of curriculum mastery and the amount of learning experienced by those who are instructed in schools. Learning achievement tests or *assessment tests* are administered to collect data on learning outcomes. Standardisation of test administration reduces bias and improves the validity of the inferences about underlying student ability and performance made from the assessment results (Greaney and Kellaghan, 2008). Standardisation is thus a feature of the best assessments and examinations.

2.2 Definitions of formative and summative assessment

According to many researchers including Lockheed and Verspoor (1991), Black & Wiliam (1998); Barber & Mourshed (2007), Braun, Kanjee, Bettinger, & Kremer (2006), Darling-Hammond & Wentworth (2010) and Clarke (2012b), effective formative assessment, although rarely specified in terms of content or techniques (Dunn & Mulvenon, 2009), is strongly associated with improved learning. Used in combination, summative and formative assessments are important features of effective learning systems globally (Darling-Hammond and Wentworth, 2010; Koretz, 2008; Conley, 2015).

The definition of formative and summative types of educational assessment varies in the literature. Harlen (2005) proposes a functional approach to differentiate between summative and formative assessment tests. According to Harlan (2005), similar assessment items or questions may be used for different purposes provided that this is done using technically defensible and transparent methods. Assessment test results may be used formatively to guide instruction within grades, classes and schools, with limited consequences for test performance by test takers, their teachers and their schools. On the other hand, summative assessment at the end of a grade or cycle is used to guide educational decisions about the post-school destination of children in the labour market and further educational opportunities. Summative assessments thus have greater consequences for low and high performance for learners, teachers and schools as the assessment results are associated with sanctions and rewards. Examinations are summative assessments which are traditionally used for selection or to control access to higher levels of education, especially in developing countries where secondary school places are restricted (Lockheed and Verspoor, 1991). The continuum proposed in Figure 2 helps in locating high-impact consequences of performance on the one side of the continuum and low-impact or no consequences for the learner and indeed, their classroom or school.

2.3 Scope, delimitations, and parameters of the study

This section deals with the delimitations, parameters and scope of the study, and the definitions of learning assessment used. Although the study is concerned with the use of sample-based standardised assessments to measure learning progress in schooling systems, it acknowledges the importance of other types of assessments which are necessary in a country's education system.

Strictly speaking, the universal and sample-based versions of ANA in South Africa are not comparable over time as the tests were not kept confidential or stable between years – the sample-based version of ANA is analysed as it has high levels of curriculum alignment and its implementation holds lessons for future assessment efforts[15]. The National Senior Certificate was also included as an exceptional case as it is sufficiently long standing albeit as a means of selection for post-school destinations (Poliah, 2014).

Various unstandardised tests were excluded including: the National Benchmark Tests, which are used to guide admission decisions in a selected number of higher education programmes of study in South Africa; assessments administered at adult learning centres and further education institutions; Provincial Common Assessment testing programmes; diagnostic assessments intended to diagnose learning difficulties at classroom- and school-level; and teacher and school performance assessments, which also fell outside the scope of this study. Some of these assessments such as the National Senior Certificate, and international assessments such as TIMSS, PIRLS and SACMEQ are, however, briefly discussed in this research

[15] Moderation: the process of establishing comparable standards for evaluating student responses to assessment tasks in order to ensure that the data are valid and reliable for the intended purposes. In schools, it involves groups of teachers looking at examples of student work, discussing the extent to which these meet the expected standard and coming to agreement on the level of attainment represented by each example. The group may consist of staff from different groups within the school, from different schools or across authorities. School-based assessment refers to assessments administered in schools and evaluated by the students' own teachers, marks from which count towards schools' or students' external/public assessment results. Typically, it involves some form of external moderation or standardisation that insures minimum comparability across different school contexts (Clarke, 2012b).

report because aspects of their implementation provide guidance and lessons for assessment system development and elaboration at country level.

2.4 Examinations and the learning assessment continuum

Detailed examination of the literature supports the notion of a continuum with *examinations* at one end for selection, usually in higher grades, or at the end of a phase of schooling with high consequences for poor performance at the individual, teacher and school level, and limited diagnostic utility as the learners will have passed into the next grade or phase of their educational career. In this continuum, *learning assessments* are located at the other end with less harsh and less public consequences for low performance and more focus on the diagnosis and remediation of instruction and learning progression within classrooms and with lower consequences in terms of public accountability.

The following extracts of literature from 1994 and 2008 support the proposal of a continuum between assessment and examination made in the next section. Greaney and Kellaghan (2008), in their widely used textbook on national learning assessment, that:

> "national learning assessments have the following purposes:
>
> - to determine how well students are learning in the education system as a whole (with respect to the aims of the curriculum, expectations of preparation for further learning and life);
> - to monitor achievements in the education system over time especially if the assessments are carried out to yield comparable information on learner achievement in different time periods;
> - to identify disparities and inequity in provision of resources and support of teaching and learning achievement - and to determine if the education system is under-serving any particular group(s) (e.g., boys/girls; language or ethnic groups; students in different types of school; with disabilities; students in different administrative/geographical locations) ;

- to identify strengths and weaknesses in students' knowledge and skills (using sample surveys);
- to identify factors associated with student achievement (using education production function methods, among others); and
- in the case of international assessments, to provide comparative data on student achievements in two or more education systems and be used to provide insight into what works in improving learning quality. "

A decade and a half earlier, Keeves (1994), in his research on the utility of national examinations, declared that:

- "Assessment results may serve as an incentive, a yardstick for schools and more broadly, for the performance of the education system; Assessment results provide the opportunity for accounting for the quality and content of schooling;

- Assessment results assist in individual progress in that they are used, if well designed, as a means of certification of the competencies of those who are assessed, and indeed, their teachers;

- International standardised assessments provide the opportunity for countries to compare and measure themselves against others; and,

- Assessments also provide the opportunity, at school level, to diagnose and remediate teaching and learning weaknesses in classrooms. However, in order to be useful the assessment system, tools and processes must be more focused on curriculum content, teacher mediation and capacitation."

Typically, countries supplement traditional examinations in a hybrid system with systemic learning assessments for diagnosis and remediation as well as sample-based systemic performance monitoring tools (Rosenkvist, 2010). Research by Gustafsson & Moloi (2011) does not find any evidence of countries' adaptation of

examinations into assessments. This supports the notion that they exist at the different extremes of a continuum with different functions and consequences.

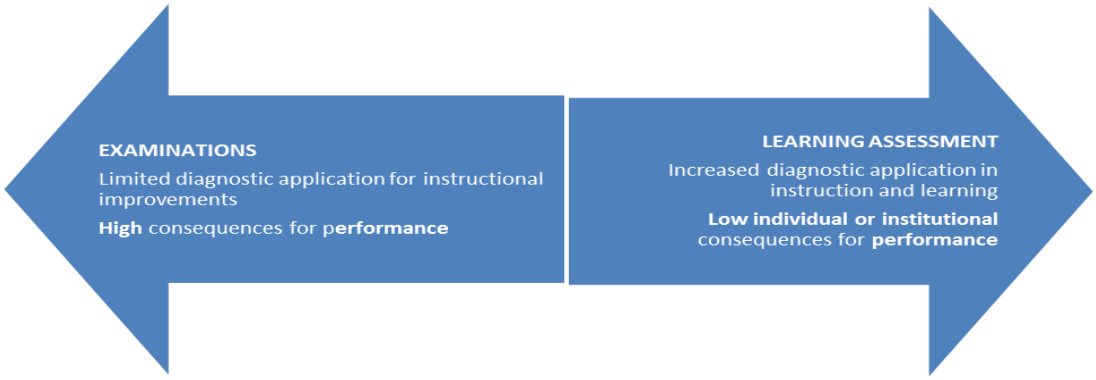


Figure 2: Examinations and learning assessment continuum

Source: Author

2.5 Use and impact of assessment data on education policy

The growing use of international learning assessments to compare and benchmark education systems has been noted in recent years by Benavot and Koselici (2015). Kamens & Mcneely (2010) identify the role of development organisations in facilitating the expansion of these assessments in support of education reform and accountability in education systems globally. Country-level responses to assessment results range from large-scale policy reform to more modest policy responses. Germany, a frequently quoted example of how radical education reform can follow assessment, introduced national curriculum standards in 2004, resulting in a widespread backwash[16] effect on post-school education following very poor PISA results in an event colloquially referred to as PISA shock in the literature[17]. Germany,

[16] According to Braun and Kanjee (2006), the "backwash effect" is the impact of assessment and particularly the use of assessment results on learning, teaching, curriculum, learning materials and education provisioning.

[17] On Page 211 of the 2010 OECD publication "Strong Performers and Successful Reformers in Education: Lessons from PISA for the United States", in the chapter entitled '*Germany: Once Weak International Standing Prompts Strong Nationwide Reforms for Rapid Improvement*' states that: "For many years, the German public and policy makers assumed that Germany had one of the world's most effective, fair and efficient school systems. It was not until 2000 that they discovered this not to be the case at all, and that in fact Germany's schools ranked below the average when compared to the PISA-participating countries. Now, ten years into the 21st century, Germany has substantially

despite its administration and delivery of the curriculum at provincial and local level, adopted national standards in 2003 and 2004, for the purposes of better monitoring of progress in learning (Grek, 2009; OECD, 2010; Rosenkvist, 2010). In contrast, PISA in Hungary ignited debate in the media but resulted in few systemic interventions. South Africa, in 2007, declined to participate in the TIMSS international assessment to concentrate on developing assessments within country according to a response from the Minister of Education, Hon N Pandor, MP. The Minister was responding to newspaper report lamenting the country not participating for that year (Mbanjwa and Kassiem, 2007). According to Clarke (2012b), deCastro (2012) and Makuwa & Maarse (2013), the results of PISA, TIMSS, PIRLS and SACMEQ have been used to initiate reform in countries as diverse as Jordan, Poland, Brazil and Namibia.

Best et al (2013), in their systematic review of the impact of national and international assessments on education policy in developing countries, confirm that these countries use assessment data in all parts of the policy cycle from monitoring, resource targeting, and advocacy. However, Brookhart (2013) notes that, despite numerous country comparisons based on rankings from international assessment results, the question of how to secure sustained learning improvement in schooling systems is still a vexed issue, excluded from much of the public debate following the results of international assessment results. Despite the limitations of international assessments for developing countries, developments such as PISA for Development (PISA-D)[18], and the introduction of the PISA for Schools Test to generate individual school report cards against PISA-aligned tests, are encouraging and can potentially improve the utility of international assessment frameworks and tests at the country level and below.

---

improved its position in the PISA league tables. This chapter explains how Germany could have so misjudged the relative quality of its education system, how it could have fallen so far from where it had been generations before, what it did to reverse its unfavourable position, and what other nations might learn from this experience. It identifies the main factors behind Germany's strong recovery as being the changes it has made to the structure of its secondary schools; the high quality of its teachers; the value of its dual system, which helps develop workplace skills in children before they leave school; and its development of common standards and curricula and the assessment and research capacity to monitor them." The OECD publication is available at http://www.oecd.org/pisa/pisaproducts/46581323.pdf:

[18] http://www.oecd.org/pisa/aboutpisa/pisa-for-development-documentation.htm

Despite the expansion of international assessment activity in recent years, the majority of the world's 144 low- and lower-middle income countries do not participate in international assessments, preferring instead to develop systemic learning assessments at the country level and to participate in regional assessments (Kamens & Benavot, 2011). This may be due to the cost of participation and/or the limited range of learning achievements covered in international assessment tests originally developed for industrialised countries. In addition, international assessments do not always discriminate between learners in many developing countries who perform at the lower end of the learner ability spectrum. International assessment results are perceived to be of policy support at macro- or system (or country) level rather than being relevant to the needs of teachers and learners in classrooms (Best et al, 2013). These considerations may explain why so few developing countries including China and India which, together, account for a third of the world's population, do not participate at country-level in large-scale international assessments such as TIMSS and PIRLS. [19]

Given that learning improvements have been shown to follow from well-developed formative assessment instruments and practices, and from mastery of the curriculum, it would be rational for these countries to focus their efforts and resources on improving systemic learning assessments and on increasing instructional improvement through classroom- and school-based assessments.

2.6 Test-based accountability, academic performance and assessment-related behaviour

In this section, the link between the use of assessment test-based accountability and academic performance is explored in the literature from different international contexts.

Research shows mixed effects of test-based accountability on learning achievement, with positive links in much of the literature although negative behaviours like cheating and student exclusion are associated with standardised testing programmes

[19] Chinese Taipei and Hong Kong do participate as discrete jurisdictions for benchmarking purposes.

where there are negative consequences for low performance (Koretz, 2008). The positive effects of test-based accountability measures in developing countries are not easy to isolate. Frequently, broader reform packages include elements of test-based accountability measures. Gustafsson & Moloi (2011) cite the example of PISA improvements in Chile and Brazil which resulted from comprehensive funding, measurement and accountability reforms which were difficult to unbundle in order to determine causality of improvements in learning outcomes.[20]

The literature on the negative effects of test-based accountability focuses on the highly visible, legislated testing culture at state level in the US (Conley, 2015) with explicit rewards and sanctions for performance (Darling-Hammond, Wilhoit & Pittenger, 2014). Antipathy in the literature is mainly expressed in relation to test-based accountability systems with a high test burden, and with public naming and shaming which is educationally unproductive at best and demotivating at worst (Koretz, 2008; Hoadley and Muller, 2016). Educational testing has always been controversial, especially in the US, where the No Child Left Behind (NCLB) legislation was introduced in 2002. Test-based accountability measures made mandatory state-wide administration of standardised tests a dominant feature of US schooling. In addition to state assessments, NCLB also required national-level test-based accountability in the form of participation in the NAEP (Hanushek and Raymond, 2005). The literature shows that a multitude of testing systems prevailed in the US with state, district and school level testing giving rise to a an inappropriately developed market in privately developed low quality assessment tests which are not always curriculum-aligned (Barton, 1999; Hoxby, 2002; Hanushek and Raymond, 2005). Despite these concerns and the complexity of valid and reliable measurement of learning progress, judicious and well-designed educational assessment testing within a comprehensive framework is a feature of many of the world's most successful education systems (Conley, 2015; Darling-Hammond et al, 2014; Rosenkvist, 2010).

[20] Brazil saw the second-largest improvements in mathematics in PISA between 2000 and 2009 (after Peru), with extensive reforms in progressive funding, accountability and measurement at national level in the last two decades (de Castro, 2012).

The antipathy to testing has been transmitted to developing countries without mediation and adjustment for context, despite the fact that many developing countries tend to have a low endowment of coherent assessment activity and capacity (UNESCO, 2015a). Section 2.7 of this chapter outlines concerns raised about test-based accountability in general and in South Africa in particular (Mweli, 2016), despite initial support for ANA by unions and civil society (SADTU, 2011; Equal Education, 2011; COSATU, 2011; SAPA, 2011). The contestation suggests that efforts to create an enabling context for broad-based assessment system reform should not be underestimated within a country (Allen, Elks, Outhred and Varly (2016).

Standardised exit examinations are linked to learning

Madaus (1988) offers an insight into the history of high stakes examinations which have a positive backwash effect on instruction, and research by Bishop (1997) and Barton (1999) linked standardized exit examination policies and practices to academic achievement – indicating the enabling role that external accountability has on country-level learning improvement, though Barton was concerned about exclusions of learners in pursuit of performance targets. The positive influence of external examinations has been put forward as a possible reason why South Africa's performance in the regional SACMEQ compared to her neighbours in the region, has been lower than expected considering the generous investment in education, of around 5% of Gross Domestic Product in the country (DBE, 2015a).

Secure high stakes examinations, with feedback, are linked to learning

Phelps (2012) confirmed that testing feedback, ranging from awareness to remediation, produces the strongest positive effect on achievement and that this effect is augmented when high stakes or consequences are linked to test performance. Phelps' recommended that standardised testing with high stakes should be more widely used with the requisite security measures. Phelps' proposals require a level of capacity and robustness of security arrangements not always available or feasible in developing countries as they need an established assessment

system and sufficient fidelity of standardised administration and processes to counteract educationally unproductive behaviour including cheating. Phelps (2012)[21] stated that:

> "Among quantitative studies, mean effect sizes range from a moderate to large (*d* ranging from 0.5 to 0.88) [22], depending on the way effects are aggregated or effect sizes are adjusted for study artefacts"

The formula used to calculate the effect size in quantitative studies could be any variation of the following formula of standardised mean difference (Cohen, 1988 cited in Phelps, 2012):

$$d = \frac{\bar{x}_t - \bar{x}_c}{S_p}$$

Where: d = Effect size, S_p = Pooled standard deviation and the means in treatment and control groups are represented by $\bar{x}_t$ *and* $\bar{x}_c$ respectively. The closer in the underlying distribution characteristics the control and treatment group are, the smaller the effect size.

Sample-based systemic assessments are less costly.

Most countries with mature education systems have universal assessments administered at school level for formative classroom-level diagnosis and feedback, supplemented by sample-based systemic assessment tests administered infrequently, in addition to international assessments for benchmarking purposes and examinations for certification and selection (Clarke, 2012a; 2012b, and Rosenkvist, 2010). Phelps' advice for secure confidential testing is relevant for

[21] According to Phelps (2012), "Several formulae are employed to calculate effect sizes, each selected to align to its relevant study design. The most frequently used is some variation of the standardized mean difference as d (Cohen, 1988 cited in Phelps, 2012). Testing with feedback produces the strongest positive effect on achievement. Adding stakes or frequency also strongly and positively affects achievement. Ninety-three percent of qualitative studies analysed also reported positive effects."

22 If both control and treatment group are very different, the effect size will be large and positive but less than 1. According to Cohen (1988), if d is less than 0.2 the effect size is small and if it is above 0.8, the effect size is substantial.

developing country contexts which are resource constrained. Despite their utility for diagnosing learning difficulties at local level, using universally administrated assessments to measure learning outcomes at country level is costly and burdensome, and difficult to standardise to reduce test result bias. This further supports the focus in this study on sample-based, confidential or secure assessment tests to reduce costs and testing burden in developing countries.

Test-based accountability systems, academic achievement and equity

With the introduction of legislation advocating state-based accountability in the US, conditions arose for a natural experiment. Many studies have shown positive links between the presence of state accountability systems before the introduction of the national accountability legislation in the US and subsequent performance on national standardised tests. Dee & Jacob (2009), in common with Carnoy & Loeb (2002), confirmed that US states with stronger test-based accountability systems exhibited higher gains in Grade 8 mathematics after national accountability laws were implemented without significantly higher exclusion, retention or other unwanted behaviour. Hanushek & Raymond (2005) found that student performance gains before and after the implementation of the NCLB legislation were substantial, though they noted concern about the proliferation and market in assessment tests, deepening of race-based inequality, and the widening assessment gaps observed as a result of the increasing concentration of minority students in particular US schools. Hanushek and Raymond, proposed policies and monitoring to counteract inequities arising from the negative behaviours associated with increased test-based accountability measures. Their concerns are relevant in developing contexts where substantial numbers of students have low levels of academic achievement. Deming, Cohodes, Jennings & Jenks (2013) showed the generally negative effects of test-based accountability on weaker performing students in their research. Cohodes et al (2013) also showed positive performance gains accruing to high-ability students in the long-run, echoing concerns about equity voiced by researchers including Barton (1999) and Hanushek & Raymond (2005).

Burgess et al (2013) showed significant reductions in school effectiveness in Wales following the decision to cease publishing school performance in standardised assessment tests after a referendum advocating a break with common reporting with England. Lower performing students were the most negatively affected by the decision.

In contrast, Rouse, Hannaway, Goldhaber, & Figlio (2007) found that in some Florida State tests schools responded to test results by improving instruction in order to secure legitimate learning gains. These schools improved their grade retention and tutoring, and paid more attention to teacher development and collaboration and to time on task. According to Bruns, Filmer & Patrinos (2011), all these are factors are linked to improved academic achievement and school quality. Education systems require monitoring of internal accountability in addition to using external accountability measures of learning performance in education systems, to minimise the unintended consequences of accountability measures in schooling systems.

Informed stakeholders linked to learning gains

Research by Hoxby (2001) found that student achievement scores improved[23] following state adoption of report card systems. Using sample-based NAEP data, she found links between the provision of stakeholder information and school level academic achievement. These were similar to the positive effects reported by Bruns et al (2011) using data in Liberia and Pakistan.

Strategic behaviour introducing bias in assessment test results

Unproductive behaviour arising from test-based accountability measures is sometimes euphemistically referred to in the literature as strategic behaviour and involves cheating, changing scores and the artificial manipulation of scores. These

[23] Statistically significant though modest reading and numeracy score increases were associated with assessment test use. At the end of ten years, nine-year-olds' reading scores were predicted as being on average 2.6 points higher and thirteen-year-olds' math scores were predicted to be 2.8 points higher.

are a global phenomenon in education systems due to the nature of examinations intended for selection (Koretz & Hamilton, 2003). Artificial score gains reduce the validity of the conclusions made about real levels of learning outcomes as they introduce bias into, and compromise the measurement of learning outcomes. It is essential to overcome or at the very least minimise the effect of improper behaviour and perverse incentives that affect the credibility of the results of assessment tests. This may be done by applying sanctions for improper behaviour, strengthening monitoring and security, documenting administrative breaches, adjusting for the effects of corrupted data and using secure confidential tests administered by an objective party or service provider.

Jacob (2005) found evidence of teachers responding strategically to test-based accountability pressures by excluding learners, or preventing learners from participating in assessments by placing them in special education classes. Earlier research by Jacob & Levitt (2003) identified behaviours such as outright cheating on behalf of students, exclusion of weaker performing learners, artificial boosting of performance on test days using dietary supplementation, and intensive test preparation.

There is some evidence to confirm that test-based accountability measures may be used to secure gains in learning achievement and test scores. However, many of these studies were carried out within years of the changes in legislation in the US. These may have yielded short-run improvements in test scores which may not have been sustained or equitably distributed in the student body in the long-run. There is limited evidence in the literature, however, to indicate the mechanisms for doing this (Brookhart, 2013).

2.7 Critique of testing: issues and considerations

This section summarises and contextualises the main critique of testing in the education literature. It also provides an overview of testing controversies and possible solutions.

Concerns about assessment efforts have mainly been expressed in the literature in relation to test burden, frequency, coverage and depth. In South Africa, Hoadley & Muller (2016) focus on concerns about testing and identify three main anti-testing arguments in their analysis of the literature and research: political objections (that testing is a tool of regulation and control and erodes trust in teachers' professionalism), pedagogical objections (that it negatively affects teaching) and conceptual objections (that assessment fragments knowledge into isolated elements). Despite the considerable opposition to testing evident in the literature, including curriculum narrowing, reduction of the breadth and depth of learning, reduced teacher autonomy (Ramatlapana & Makonye (2012)), demotivation of teachers and inappropriate curriculum management practices exacerbated by testing, building on the work of Singh, Martsin & Glasswell (2015) cited in Hoadley & Muller (2016) who conclude that assessment can produce benefits provided that teacher confidence and capacity in assessment practice are nurtured.

Concerns about curriculum narrowing are countered by the observation that, in addition to the traditional assessment of language, numeracy and life skills subjects, subjects such as science, social science and foreign languages are now more common in national assessments as issues of global competitiveness influence skills development and supply in countries (Benavot & Koseleci, 2015). For example, NAEP also includes assessments in the Arts, Civics, Geography, US and World History as well as Economics, while regional assessments such as the fourth cycle of SACMEQ included life skills modules with elements relating to knowledge of TB, HIV and AIDS.

Darling-Hammond, Wilhoit & Pittenger (2014) confirm the systemic benefits of well-designed standardised educational assessment programmes as part of an effective educational assessment system. These researchers counter the criticism that teachers will teach to the test or that tests encourage cheating. They suggest that, in the absence of teacher assessment capacity or an effective formative assessment programme, well-designed standardised tests can help to align classroom learning and instruction and to communicate the curriculum to new or inexperienced teachers.

Research findings about the optimum balance of consequences or stakes to be used in educational assessment are mixed. Linn (2001) proposed reducing testing stakes to reduce corruption, gaming and exclusion associated with high stakes assessment. Phelps (2006) argued that instructional narrowing, constrained instructional innovation and creativity will be minimised with a well-understood and credible framework and system for formative and summative assessment.

Critique of assessment transcends country boundaries. Despite initial enthusiasm with the annual national assessment following the President's announcement of the programme (Zuma, 2011), concerns over the years from teacher unions and teacher associations have ranged from concern about the test burden to ideological opposition to the misuse of PISA results for ranking in order to pronounce on the quality of education systems driven by an assessment agenda which is set by industrialised countries (Education International, 2013). In South Africa, as late as 2014, the largest teacher union in the country declared its opposition to over-emphasis on assessments and tests in the education system with support from political groupings and other unions (DA, 2014). The SADTU National Congress of 5 October 2014 re-affirmed the 2013 National General Council (NGC) resolution:

> "That ANA should remain a systemic evaluation with clear time frames that would allow for prompt feedback to be given to schools before the results are publicized followed by meaningful intervention programmes; that ANA should not be abused to label teachers and schools, thereby demoralising and de-professionalising them; and, that ANA should be reviewed as an annual assessment as of 2015, and be substituted by a 3 year cycle of assessment."

The union also proposed a review of processes, tools and synergy between all existing assessment tools (SADTU, 2014). Essentially, SADTU's concern related to the costs and burdens of the universal testing programme, although the organisation supported sample based systemic assessment (which would have low consequences for individual schools and teachers). Many of these concerns are considered in the recommendations arising from this study more so as they represent the perceptions

of practitioners in the education system, and affect the enabling context of assessment reform.

Concerns about construct validity have also been expressed in South Africa. Gravett and Henning (2014) identified the lack of a scientific construct used in test design and limitations in the diagnostic utility of the ANA. They recommended, better item and test design and research into context-appropriate assessment tools, items and instruments for use in classrooms to supplement the measurement of learning progress at the system or country level. Rather than dismissing critiques of testing, this study used the concerns raised to inform the feasibility of the policy and research recommendations made.

2.8 Frameworks for analysing effective student assessment systems

This section presents various frameworks for analysing assessment systems in education, which vary from the purely descriptive to more analytical efforts in the literature.

Braun & Kanjee (2006), in their reflection on the use of educational assessment in improving education (systems), focused on describing the features, roles, utility and use of assessment in developing countries. Their work was informed by a series of technical engagements in 2002 on the role of educational assessment in achieving universal basic education. Their approach highlights the role assessment can play in extending access, quality, equity and efficiency in education systems in the developing world using case studies and findings from the literature. They state:

> "This paper provides a framework for conceptualizing the various roles assessment plays in education.....it suggests how assessment practices and systems can generate relevant and timely information for the improvement of education systems, presents case studies of a number of nations, describes some international efforts, and proposes next steps."
>
> "Certainly, assessment data, when appropriately aggregated, can be an important component of a broader educational indicator system. This paper, however, does not treat the use of assessment for such administrative purposes as the evaluation of teachers, principals, or schools. ….In this paper,

> we first propose a framework through which we conceptualize the role of assessment for improving access, quality, efficiency, and equity (AQEE) within the education system."

Braun and Kanjee acknowledge the need for assessment information which provides opportunities for improvement in any other attributes they identify as being important for a credible education system. They identify several considerations for countries embarking on the development of effective assessment systems including: conducive policy context and institutional arrangements; strengthened assessment practice, experience and personnel capacity for assessment of different types; linkages between different kinds of assessment and educational activities in the education system; and the provision and availability of appropriate physical, human and technical resources and specialists. They also identified the need for educational improvement in the AQEE attributes as the desired outcome of any assessment efforts at country level. Their research confirms the importance of the technical quality and standardisation of aspects of assessment implementation including appropriate sampling, inclusiveness, processing, documentation, analysis, reporting, and disaggregation of results – all of which should be understood and based on credible evidence.

Ferrer (2006) developed an analytical framework for national assessment systems; this is shown in Table 1 below. In the Partnership for Educational Revitalization in the Americas (PREAL) project, Ferrer documented the progress, profile and development of Latin American and Caribbean countries' assessment systems at national and subnational level, based on 60 interviews and on documentary and literature review in 19 countries and 5 subnational systems.

The OECD framework for analysing evaluation and assessment in country-level education systems focuses on student assessment, teacher appraisal, school evaluation, school administration and leadership evaluation and system evaluation. Within each of these dimensions, the main features covered in the country analyses include reflections on governance, design, capacity building, the use of assessment results and implementation strategies. The framework was developed from information gleaned from 28 OECD countries about the main features of their

national and sub-national educational assessment and evaluation systems (OECD, 2013). The review identified issues related to governance, procedures, capacity, reporting and use of reports and results to inform education system improvement. Thereafter the review identified the policy options in Box 1 below as a means to strengthening the use of education system evaluation and assessment for system improvement.

Box 1: Education System Evaluation: Informing Policies for System Improvement

Governance: Ensure a broad concept of education system evaluation within the evaluation and assessment framework; Ensure policy making is informed by high quality measures, but not driven by their availability; and, situate education system evaluation in the broader context of public sector performance requirements

Procedures: Develop a national education indicator framework; Design a national strategy to monitor student learning standards; Ensure the collection of qualitative information on the education system; Assure the monitoring of changes over time and progress of particular student cohorts; and, ensure collection of adequate contextual information to effectively monitor equity.

Capacity: Establish and secure capacity for education system evaluation; Promote the development of evaluation capacity at the local authority level; and, ensure objectivity and credibility in education system evaluation activities

Reporting and use of results: Strengthen analysis of education system evaluation results for planning and policy development; Communicate key results of education system evaluation to stakeholders; and. support feedback for local monitoring

Source: OECD, 2013. Synergies for Better Learning: An International Perspective on Evaluation and Assessment. Available at *www.oecd.org/edu/evaluationpolicy*.

Clarke's framework (Clarke, 2012b) and the evaluation matrix (SABER, 2014) provide a means of evaluating four different types of educational assessment programmes based on ratings of the level of development of each dimension and sub-dimension of large scale assessment. The framework also provides the basis of an evaluation matrix (SABER, 2014) against which assessment programmes may be compared. Her framework, arising from country experiences of policy makers and assessment practitioners, consists of three dimensions of an effective student assessment system: the enabling context, system alignment and assessment quality. Together,

these three dimensions accommodate the attributes identified by the frameworks described by Ferrer, Braun and Kanjee, and even the OECD This alignment makes Clarke's framework the most appropriate for use in this study, and it can also be used in combination with the rubrics developed from the framework which have been tested in different countries in the SABER project (SABER, 2014).

Table 1. Analytical themes used in developing an overview of national education systems (reproduced from Ferrer, 2006)

Analytical category	Themes
Institutional framework	Adequacy to political framework and technical capabilities Stability Financing and administrative autonomy Human resources Autonomy and capacity to disseminate results Transparency
Curriculum and standards	Availability, adequacy and use of national curricular frameworks to design assessment instruments Development and validation of standards consistent with the prevailing curriculum
Instruments	Solid and explicit conceptual framework for drawing up reference matrices Validation of reference matrices and instruments Types of items Cultural and linguistic adequacy (especially where bilingual and intercultural education programs should or do exist) Sample-based or census-based coverage consistent with the aims of the assessment Study of the in-school and out-of-school context for analysis of performance-related factors
Reports	Coherence between types of reports and expected uses (curricular development, pedagogy, targeting support, teacher training, selection of students and so on) Adequacy of reports for different audiences (clarity, guides to interpretation, sensitization and so on) Information on associated factors and value-added models
Dissemination and uses	Delivery: time frames; scope; regularity Impact: school use; policymaking/program design; political accountability High stakes: schools; teachers; students
Subnational systems	Main differences from national systems: Standards Sampling (or censal) coverage Participation of local actors Time frames and formats for delivering results Use for pedagogical improvement
International tests	Development of technical capabilities Dissemination of results and impact on public opinion Specific use of the results

Rather than assuming that Clarke's framework holds, evidence is presented in this section of the research report to support the different dimensions of the original and modified version of Clarke's framework. Although Clarke's framework applies to all types of educational assessments (classroom-based assessment, examinations and international and national systemic learning assessments), this study focuses on standardised systemic sample-based learning assessments which Clarke called National Large Scale Assessments (NLSAs).[24]

The modifications made to Clarke's original framework, and the evaluation matrix developed by SABER, 2014 arising from this framework, include specific sub-dimensions of system alignment and assessment quality which should be explicitly included in any assessment programme. The modifications proposed are specific enough to contribute to the system alignment and assessment quality dimensions of Clarke's framework (2012b) and the SABER rubric for analysing assessments. The modifications include consideration of the general understanding of the assessment and links to the rest of the assessment and education development context, and in relation to assessment quality, the modifications include the consideration of: security and confidentiality of assessment; credibility of the standardisation of the learning assessment programme; item development; test design; technical sampling integrity and design; specific analytical considerations in testing and reporting; measurement of test score growth; and the development of performance standards and achievement levels in tracking learning progress at country level. The modifications were used in the analysis of South Africa's systemic learning assessment programmes in Chapter 3, using an evaluation matrix based on the modifications to the SABER rubric.

Table 2 below shows the stages of development of a student assessment system from emerging, established and advanced levels of development of the dimensions

[24] Firstly, there must be political leadership, public understanding and sustainable technical capacity to ensure an enabling context with technical integrity. Secondly, system alignment must include linkages between the curriculum and instruction, and these must be enhanced using assessment information. Thirdly, the components of the quality of the assessment must be technically credible from the design and development of assessment instruments through to the administration of instruments, sampling, data collection, analysis, interpretation and reporting of the assessment results.

according to Clarke (2012b). Table 3 illustrates the classification of four different types of learning assessment making up a typical education assessment system, and Table 4 focuses only on systemic learning assessments (what Clarke refers to as National Large Scale Assessments) intended to monitor progress in learning outcomes at the country level.

Table 2: The development stages of a student assessment system (from Clarke, 2012b)

	EMERGING (on the way to meeting minimum standard)	**ESTABLISHED (acceptable minimum standard)**	**ADVANCED (best practice)**
Enabling context	Limited policy framework, weak leadership and public engagement, limited trained staff and high turnover, irregular and unpredictable funding, unstable institutional and structural arrangements.	Clear policy and legislative framework or guidelines, Strong leadership and public engagement, trained staff with low turnover, stable institutional arrangements with clear reporting lines and predictable funding.	As for established, with a strong innovative focus on linkages between different types of assessment including teacher evaluation, role of teachers, school based and classroom assessments. Characterised by innovation, research and evidence-based practices.
System alignment	Assessments not fully aligned with learning goals, standards and curriculum. Assessments not aligned with teacher training pre-service and in-service development.	Assessment fully aligned with learning goals, standards and curriculum. Assessments aligned with teacher training pre-service and in-service development.	
Assessment quality	Limited awareness of application of technical and professional standards for ensuring assessment quality.	Some awareness and application of technical and professional standards for ensuring assessment quality is effective and is credibly used.	

Note: The latent level is omitted because it represents the absence of any assessment activity.

Table 3 : Development levels for different types of assessments (reproduced from Clarke 2012b Annexures)

ASSESSMENT TYPE	LATENT (absence of or deviation from the attribute)	EMERGING (on the way to meeting minimum standard)	ESTABLISHED (acceptable minimum standard)	ADVANCED (best practice)
Classroom assessment	There is no system-wide institutional capacity to support and ensure the quality of classroom assessment practice.	There is weak system-wide institutional capacity to support and ensure the quality of classroom assessment practice.	There is sufficient system-wide institutional capacity to support and ensure the quality of classroom assessment practice.	There is strong system-wide institutional capacity to support and ensure the quality of classroom assessment practice.
Examination	There is no standardized examination in place for key decisions.	There is a partially stable standardized examination in place and a need to develop institutional capacity to run the examination. The examination typically is of poor quality and is perceived as unfair or corrupt.	There is a stable standardized examination in place. There is institutional capacity and some limited mechanisms to monitor it. The examination is of acceptable quality and is perceived as fair for most students and free from corruption.	There is a stable standardized examination in place and institutional capacity and strong mechanisms to monitor it. The examination is of high quality and is perceived as fair and free from corruption.
National (or system level) large-scale assessment (NLSA)	There is no NLSA in place.	There is an unstable NLSA in place and a need to develop institutional capacity to run the NLSA. Assessment quality and impact are weak.	There is a stable NLSA in place. There is institutional capacity and some limited mechanisms to monitor it. The NLSA is of moderate quality and its information is disseminated but is not always used in effective ways.	There is a stable NLSA in place and institutional capacity and strong mechanisms to monitor it. The NLSA is of high quality and its information is effectively used to improve education.
International large-scale assessment (ILSA)	There is no history of participation in an ILSA or plans to participate in one.	Participation in an ILSA has been initiated but there still is a need to develop institutional capacity to carry out the ILSA.	There is more or less stable participation in an ILSA. There is institutional capacity to carry out the ILSA. The information from the ILSA is disseminated but is not always used in effective ways.	There is stable participation in an ILSA and institutional capacity to run the ILSA. The information from the ILSA is effectively used to improve education.

Table 4: Development levels for National Large Scale Assessments (NLSA) (reproduced from Clarke, 2012b Annexures)

	NLSA LEVEL OF DEVELOPMENT			
National Large Scale Assessment (NLSA) dimension	**LATENT (absence of or deviation from the attribute)**	**EMERGING (on the way to meeting minimum standard)**	**ESTABLISHED (acceptable minimum standard)**	**ADVANCED (best practice)**
Enabling context	No NLSA has taken place and there is no plan for NLSA. Stakeholder groups oppose the NLSA and no funding exists. No office or team exists.	NLSA has been operating with informal policy documents on an irregular basis with some funding to cover core operations but no funding for research, development and innovation. Political considerations hamper technical considerations on the assessment frequently. Accountability may be unclear and the NLSA unit still has capacity constraints.	A stable NLSA is held with predictable regularity. Formal policy documents on intentions, goals and operations are available and regular sustained funding to cover all aspects of operation is available. The NLSA unit and staff are adequately capacitated and report to a recognised body. Country offers some limited opportunities for preparation for work opportunities on the NLSA.	There is a detailed plan for the NLSA for the medium and long term, and funding covers research and development activities which are planned and executed competently by adequate numbers of staff. The country/system offers wide range of opportunities for preparation for work opportunities on the NLSA.
System alignment	NLSA alignment with learning and curriculum standards is not clear. NLSA measurement is questioned by the majority of stakeholders and the confirmation and evidence that the NLSA measures the right constructs is not always clear. No courses on the NLSA.	NLSA alignment is clearly linked to the curriculum and ad hoc reviews are done to ensure measurement of intended constructs. Occasional courses on the NLSA.	NLSA measures performance against learning standards clearly and there are regular internal reviews and clear reporting of alignment of the NLSA with national curriculum objectives. Regular courses are offered on aspects of the NLSA.	NLSA measurement is accepted by most stakeholder groupings. High quality courses and workshops on the NLSA offered on a regular basis.

	NLSA LEVEL OF DEVELOPMENT			
National Large Scale Assessment (NLSA) dimension	**LATENT (absence of or deviation from the attribute)**	**EMERGING (on the way to meeting minimum standard)**	**ESTABLISHED (acceptable minimum standard)**	**ADVANCED (best practice)**
Assessment quality	No technical report on the quality NLSA exists. No attempts at inclusion of all student groups are made. No results are disseminated; results are not used according to the technical characteristics of the report. No method for monitoring the use and consequences of NLSA exist.	Some technical documentation exists on aspects of the NLSA are contained in reports. Results are poorly disseminated and rarely used.	At least one option for inclusion is offered to student groups. There is a comprehensive technical report with restricted circulation, NLSA results are consistently used in a way that is consistent with their original purpose and technical characteristics by some stakeholders. Some methods for monitoring the use and consequences of NLSA exist.	Different options for all student groups are offered for inclusion in the assessment with a comprehensive high-quality technical report available to the general public. NLSA results are consistently used in a way that is consistent with their original purpose and technical characteristics by all stakeholders. Variety of methods used for monitoring the consequences of NLSA exists.

2.9 Evidence for the technical components, dimensions and standards of effective standardised systemic learning assessments

This section provides the evidence for the dimensions used in the original and modified version of Clarke's framework that may be used to classify the levels of development of educational assessment systems in relation to three dimensions of an effective student assessment system: the enabling context, system alignment and assessment quality. Some of the modifications to the dimensions and sub-dimensions of the framework may have been implicit in the original framework, however they are presented her in detail as they provide information which may be used to strengthen practice and policy in implementing assessments for tracking learning progress at country level. The evidence presented provides the opportunity to reflect on the best practice and in addition to evidence from the literature, techniques and methods used in international standardised assessment programmes which are acknowledged as exemplary in the literature by researchers including Braun & Kanjee (2006), Wagner (2010) and Cresswell, Schwantner & Waters (2015).

The modifications to Clarke's original framework and the SABER rubric were effected, using findings in the literature and technical and other documentation on best practice, to improve the technical and non-technical dimensions relating to system alignment in terms of general understanding of the assessment and specifications to sub-dimensions of assessment quality relating to standardisation, security and reporting, and item development and design. The analysis of systemic learning assessment efforts in South Africa, using the modified evaluation matrix derived from best practice and the literature and the SABER rubric, is presented in Chapter 3. Chapter 4 presents a narrative overview of systemic learning assessments used for such measurement at the country level in South Africa.

2.9.1 Enabling conditions and context

This section deals with the evidence for the requisite institutional and policy context, social validation and support, resourcing, capacity, funding and sustainability of the assessment programme for measuring progress in learning at country level. These are explored in all frameworks examined (Braun and Kanjee, 2006; Ferrer, 2006; OECD, 2013). Clarke (2012b) indicates the need to generate responses to the following questions in order to classify the level of development of the enabling context: What is the extent of the enabling institutional and policy conditions required for the assessment? Is there an overall policy framework (including intent, objectives and goals)? What is the extent of public engagement and stakeholder support of the assessment? Is there funding of the assessments? Is there integrity of organizational structure? Are there sufficient human resources for assessment?

Institutional and policy context (and alignment).

In advanced systems, a coherent institutional policy and support context for systemic assessment should be developed, implemented and communicated along with the institutional framework, goals and priorities for education assessment and education reform more generally (Ferrer, 2006). Communication should repeatedly reinforce the intended use of assessment test performance in monitoring progress, in staff development, in supporting teaching and in feedback to parents. All communication material should relay consistent messages on the real uses and consequences attached to test performance, with clear implications for teachers, learners, schools and parents. The assessments must be fully and comprehensively understood by learners, teachers, education officials, parents and other parties interested in the education enterprise. Communication must include technical expertise and specialist advice.

In South Africa, Cartwright (2013) noted the levels of confusion, even among departmental officials, on the role and functions of the new Annual National

Assessment (ANA). In addition, he noted to the lack of involvement of curriculum and psychometric specialists in the development of the ANA, and the resulting misalignment between assessment outcomes and impact and instruction. This situation is similar to that which prevailed in the US a decade ago (Hanushek & Raymond, 2005). It could be argued that the difficulties experienced with ANA were due to a lack of shared understanding about what it should and should not be used for, in relation to the education system. In the absence of a comprehensive tool for school performance monitoring at the country level, and despite the lack of test comparability in the sample-based and universal ANA between years, performance in universal ANA in different years was used inappropriately to sanction or reward schools for accountability purposes. This emerged in meetings with union representatives in 2016 in response to union opposition to ANA and is evident in the representations of the South African Democratic Teachers Union (SADTU), the largest teacher union South Africa.

Social validation

Eliminating linguistic and cultural exclusions in assessment test instruments are more appropriately addressed in section 2.9.3 below in relation to inclusive assessment instruments and items. Social validation refers to a general understanding and agreement on the utility, processes and coherence of the implementation of different types of assessments in an education system. Ferrer (2006) refers to this and Clarke (2012b) articulates it as social validation.

Sustainable funding

Research by UNESCO (2015) and Lockheed (2008) confirms the benefits of investing in measuring learning at country level over time. Hoxby (2002) agreed and estimated the costs of systemic assessment at approximately 0.07 percent of the education budget in states in the US and Latin America or less than ten dollars per student. International assessments such as TIMSS and PIRLS cost almost triple this amount per student according to World Bank estimates at the time (Lockheed, 2008), due to the costs of specialists and expenses associated with scoring and data processing in

such studies. According to the Department of Basic Education (2016)[25], South Africa's allocation of funds for annual assessment activities was approximately R167 million in 2015/16, a figure approximating Hoxby's guideline of 0.07% of the national budget.

Integrity of organisational structure, capacity and human resources

Funding, administrative and technical capacity and effective organisational arrangements determine the sustainability of assessment systems. Industrialised countries in the OECD typically have advanced systemic assessments typically with independent or semi-autonomous agencies with specific functions to report on the equity and quality of education at system level (Rosenkvist, 2010). Such agencies are separate from the curriculum provisioning function of the Ministries of Education which tend to focus on classroom-based and school-based assessment, and draw on independent curriculum and psychometric expertise in test development and analysis. Brazil's Instituto Nacional de Estudos e Pesquisas Educacionais Anisio Teikeira (INEP) is an example of such a federal student assessment institute which was consciously repositioned and resourced to effect test-based accountability programmes in that country (de Castro, 2012). Kenya established the National Assessment Centre in 2006 to carry out systemic assessment research and developments to drive education assessment innovation (Wasanga and Ndege, 2012).

Highlighting the importance of capacity in effective assessment programmes, Lockheed (2008) identified the weaknesses in technical capacity in developing countries especially in data manipulation, analysis and psychometric techniques. In addition, in some countries, staff associated with examinations also get involved in carrying out the assessment function without fully understanding the technical details and protocols for carrying out standardised learning assessment (Chakwera, Khembo and Sireci, 2004). Until 2015, South Africa's ANA was administered by the examinations and assessment unit of the National and Provincial Departments of Basic Education (DBE) although external assistance was sought on the sample-based

[25] DBE personal communication with K Matjiu, Financial Services unit, Department of Basic Education, South Africa. 6 June 2016

systemic assessment and various associated analytical and reporting activities[26] (Mweli, 2016). Operational efficiency may have been tested by these arrangements and complaints of testing overload may have aggravated by the dual roles these staff and teachers have, in addition to their quarterly provincial assessments, had to play as test administrators, scorers and moderators in the cycles of Annual National Assessment implementation.

2.9.2 System alignment

This section presents the evidence for the system alignment dimensions of Clarke's original framework in relation to how aligned all aspects of a country's education system are to support effective systemic learning assessment. Clarke's research indicates the need to answer the following questions in order to classify the level of system alignment to support effective systemic learning assessment in a country's assessment system: Is the assessment aligned with the other components of the education system: curriculum and learning goals and standards, teacher development? Are teachers acquainted with the elements of the assessment: student population coverage, domain coverage and depth, usefulness in relation to nationally-agreed and stakeholder-agreed goals and priorities for learning?

Learning assessments aligned to the curriculum

Systemic learning assessment programme need to generate information on learners' acquired knowledge in relation to existing curriculum standards (Loveless, 2005). Countries such as Germany, post-Apartheid South Africa and Sweden have adopted national curriculum standards (Mullis, Martin, Minnich, Tanco, Arora, Centurino, & Castle (2012).

In the US, despite federal arrangements, the de facto national standards are set through the NAEP and the emerging Common Core Standards for curriculum which

[26] Discussion in October 2016 with Dr R Poliah, Chief Director responsible for assessment and examinations, National Department of Basic Education, Pretoria.

have been voluntarily adopted by the majority of states in support of learning improvement in recent years. [27.]

Although there is a trade-off between coverage of country-specific curricula and comparability in international assessments (Schiefelbein and McGinn, 2008), attempts are made to accommodate country-specific questions in surveys such as TIMSS and PIRLS. However, the match is rarely satisfactory as international assessment tests are not specifically aligned to curriculum standards at the country level. South Africa's review of TIMSS science items showed, for instance, that only a fifth of the items in TIMSS 2011 matched the national science curriculum of grade 7 students while half of the items matched the grade 8 science curriculum (Howie & Hughes, 2000) although this has been reported to have improved to over 80% in many areas of the curriculum in the TIMSS 2015 exercise (Reddy et al, 2016). The regional SACMEQ assessment involved a documented country curriculum alignment process as part of implementation prior to 2005 (Ross, Saito, Dolata, Ikeda, Zuze, Murimba, & Griffin, 2005). There is no documentary evidence to confirm how this process was implemented again after the second SACMEQ cycles. There are, however, verbal assurances from South Africa's SACMEQ National Research Coordinator to confirm this alignment for SACMEQ tests[28].

Because of the emergent nature of the recently-introduced ANAs, this condition did not hold in the sample-based ANA between 2011 and 2013 and the different ANA papers were used with different assessment frameworks for the different items in the two years examined (SAB & T Deloitte, 2013).

[27] The US NAEP exists in a federal system which has no national curriculum standards. It has had to develop an assessment framework based on desirable knowledge, skills and competencies arising from a governance structure at national level called the National Assessment Governing Board (NAGB) and have selected the subjects assessed by NAEP. NAGB oversees the creation of the frameworks that underlie the assessments and the specifications that guide the development of the assessment instruments. The framework for each subject area is determined through a consensus process that involves teachers, curriculum specialists, subject matter specialists, school administrators, parents, and members of the public.

[28] Personal communication with Dr M Chetty, Acting Director, Department of Basic Education, 2016.

Teachers' understanding of learning assessment

General understanding of teachers and other personnel of assessment cycles, components and functions is an important factor in high performing education systems (Darling-Hammond & Wentworth, 2010). Three decades of education research studies show the link between insufficient instruction, insufficient opportunity to learn, low levels of classroom feedback, weak instructional practice and poor learning outcomes. Research carried out by Lockheed & Verspoor (1991) and Black & Wiliam (1998) confirms this link. South African researchers have consistently identified poor assessment practice, techniques, capacity and feedback in classrooms as systemic weaknesses.

Teachers are recognised as critical in the measurement of learning and assessment (Taylor & Vinjevold (1999); Umalusi (2004); DoE (2003b); DoE (2005); Van der Berg & Shepherd (2010); DBE (2010a); Shepherd (2011); Carnoy Chisholm & Chilisa (2012); Carnoy, Chisholm & Baloyi, 2008; Taylor, S. (2011); Hoadley (2012)). Countries such as Uruguay have worked to ensure alignment between instruction and assessment in the education system over a period of two decades (Ravela, 2005). Brazil took a similar amount of time to institutionalise the link between assessment reform and accountability reforms at school and system level. It carried out extensive realignment of information systems for instructional support, curriculum coverage and tools for technically valid measurement of learning progress (de Castro, 2012).

In South Africa, the development of general and specialised assessment capacity among education professionals has not kept up with curriculum reform. For example, short professional development programmes for teachers offered by higher education institutions were contained in a short course catalogue on the website of the DBE in 2014. The catalogue indicated that, out of 318 education courses at South African higher education institutions, only 65 were assessment related. These accounted for just 3% of the total credits (19 604) of all education courses available for upgrading teachers in that year.[29] This situation is not conducive to the establishment of an assessment system. Teacher assessment

[29] Calculated from data provided by the South African Department of Basic Education in the Higher Education Institution short course catalogue (downloaded 26 May 2014 from www.education.gov.za)

capacity and ability should be prioritised for development in service and pre-service offerings to teachers for their skills development, as articulated in the basic education sector plan (DBE, 2015a).

2.9.3 Assessment quality

This section presents the evidence for the assessment quality dimensions of Clarke's original framework, in relation to the how appropriate the country's education system is in terms of the quality of systemic assessment practices and processes. Clarke's research indicates the need to answer the following questions in order to classify the enabling context for systemic learning assessment in a country's assessment system: Are there mechanisms in place to ensure the inclusiveness and quality of the assessment? Are the assessments used properly? Clarke suggests that assessment quality should include a reflection on adequate technical documentation on the assessment, effective dissemination and appropriate use of results by all stakeholders which is consistent with the technical characteristics of the assessment.

The modification of Clarke's framework and the SABER rubric carried out in this emphasises where governments in developing countries such as South Africa should invest in order to more meaningfully measure the progress of learning outcomes. For example, supporting test and item development using psychometric means for to ensure test difficulty stability and equivalence would improve the low ratings attached to test design in all examinations and assessments in the system, provided that these efforts are rigorously carried out. Better test design, item development, calibration and piloting, and comparability are crucial in sustaining future systemic learning assessment efforts in South Africa. The importance of technical support in the areas of weakness identified in the analysis in this study cannot be overstated. The modifications recommended in this study may be incorporated into the assessment quality dimension of Clarke's framework as they provide further specification of areas which, if addressed, can contribute to strengthening the assessment system in South Africa, and more broadly, in the assessment systems of other developing countries in order to better track progress in learning outcomes.

Inclusiveness of the assessment

The importance of full participation of the learner cohort to ensure the validity of the assessment results as accurate reflections of student ability in the system. Detecting patterns of exclusion is important for ensuring inclusion in educational assessments (Clare, 2012). For example, routine comparisons are made in NAEP to detect exclusion spikes at school level before and after the administration of tests. Strict protocols exist for international assessments such as TIMSS and PIRLS for accommodating learners with special assessment needs, and for excluding learners whose disabilities prevent them from participating in assessments according to Bandeira de Mello, Blankenship, & McLaughlin (2009). TIMSS sampling design allows the exclusion, at school level, of mentally disabled learners or those diagnosed with physical disabilities, language barriers or other difficulties which do not allow their participation in the tests (Joncas and Foy, 2011).[30] Clearly such learner populations require in-depth and dedicated learning measurement programmes directed specifically at them to supplement the more general systemic learning assessment efforts in a country as in the Systemic Evaluation carried out in special schools among disabled learners within special schools in 2003 (Department of Education, 2003c).

Technical documentation and knowledge management

Documentation assists in codifying methodologies, analyses and research processes so that they can be replicated or scrutinised. The levels and detail of technical documentation in international assessments such as TIMSS, PIRLS, PISA and national assessments such as the NAEP are taken as the 'gold standards' in this respect according to Cresswell, Schwantner, & Waters (2015). Documentation is freely available on the internet to cover all aspects of instrument development, sample design, data collection, processing, scoring, data management, analysis and scaling in addition to decision-making about survey content. Many of these statistical and

[30] Although this seems unfair, allowing excessive time for completion of standardised assessments may overstate the performance of disabled learners, and introduce bias into estimates of learner ability in the population. This in turn, could compromise targeted support interventions for this set of learners in addition to other decisions about relative provisioning in the education system.

technical methodologies used in international assessments were adapted from the NAEP (Foy, Gallia & Li, 2007), as can be seen from the documentation.

In South Africa, the instruments used in the Monitoring Learning Assessment (MLA) in 1999 are not in the public domain. At regional level, technical documentation on some aspects of implementation including the standardisation processes are not available in updated form on the implementation of the Southern and Eastern African Consortium for Monitoring Education Quality study (SACMEQ) in the years following 2007. Other than in a research paper on sampling methodology in 2005 (Ross et al, 2005), comprehensive documentation on SACMEQ implementation is not publicly available after this. This leads to concerns about the reliability of the reporting and the level of standardisation, although the SACMEQ Coordinating Centre verbally confirmed late in October 2016 that this technical documentation will be available in the near future[31]. Developing countries are not unique in having documentation problems. Research by Jerrim (2013) showed that counter-intuitive PISA trends observed in the UK arose from a change in methodology not reflected in the UK's PISA documentation at the time and Kreiner & Christensen (2014) reflect their discontent with the documentation and resultant ranking of scaling methodologies in PISA. This highlights the importance of country-level documentation. PISA results for the country indicated a decline in performance over the years and this was used, incorrectly, to score political points by opposition parties (Young, 2011). TIMSS trends were opposite to those reported in the PISA results over the same period. Jerrim's research indicated that an undocumented methodology change was responsible for the observed PISA trend.

Security and confidentiality concerns

Advanced education systems typically use a combination of confidential sample-based systemic tests, locally developed tests for instructional improvement and classroom-based formative assessment – the latter need not be confidential as they are used to diagnose and remediate learning and instructional difficulties in the classroom (Rosenkvist, 2010) while the former need to be standardised. There is

[31] Discussion with Ms T Masalila, Head of the SACMEQ Coordinating Centre, Botswana on 20 October 2016.

always a balance between the validity reduction offered by standardised tests kept confidential over time, and the need to reduce test exposure by repeatedly administering the same test over time. TIMSS and PIRLS as well as the NAEP refresh a portion of the items in each cycle of tests.

Independent service providers may help in maximizing confidentiality and this approach is used in NAEP test administration (NCES, 2015). South Africa's Western Cape Province runs its own sub-national Systemic Evaluation assessment programme using independent service providers to annually administer confidential tests to all learners in specific grades in schools[32] as the national Department of Education did in the Systemic Evaluation survey in 2001, 2004 and 2007. Independent test administration service providers for the Western Cape are responsible for collecting the test instruments, eventual data capturing, scoring, standardisation and analysis within a secure environment. A small proportion of test items are refreshed for every assessment cycle, and for the sake of continuity within the assessment programme, service provider appointments are typically made for more than one cycle according to the officials managing the programme.

Standardisation of learning assessments: reducing bias.

Standardisation in educational assessment requires constant effort and attention to avoid bias in results. Standardisation requires that test administration and scoring are predetermined in a standard manner so that results can be attributed to differences in student ability and not to differences in administration. Lockheed (2008) proposed standardisation in test design, analysis and processing in addition to administrative procedures although she conceded that few developing countries have the technical capacity to apply, document and assure standardisation techniques consistently during test administration, development, analysis and reporting. Standardised tests may have any format but are frequently in multiple choice question formats, typically scored by computer, leading to the incorrect assumption that all standardised tests have a MCQ format. If carried out by human judgement only, subjective judgement can arise in the selection of questions,

32 Personal communication with Dr Andile Siyengo, Director Research responsible for systemic testing in the Western Cape Education Department, February 2016.

phrasing of questions and setting passing scores and this can introduce bias in test results. Test developers typically discard a large number of items answered correctly by too many or too few students including after field trialing, and following psychometric and other measurements.

Standardisation of assessment test items also ensures validity, that reliable tests are used and that the results generated are not biased.[33] Standardisation should also cover monitoring, reporting and analytical processes. In South Africa, the appointment of independent service providers in the administration of the sample-based ANA has assisted in standardisation of test administration. The lack of comparability of the ANA tests between years has dogged the sample-based ANA results and has been reported by the Department of Basic Education itself (DBE, 2012b) and other researchers (Cartwright, 2013; Spaull, 2013b). Other issues relating to standardisation include implementation, monitoring or reporting fidelity. For example, the official report on the 2013 Sample-based Verification ANA states that the emphasis in 2011 on scoring fidelity by teachers was not carried over from 2011 despite the obvious utility of monitoring these trends in this indicator. In addition, the sampling for 2011 excluded small schools but included these schools in proportional population sampling used in the 2013 exercise (DBE, 2011a; SAB & T Deloitte, 2013). Such issues may be refined in later iterations of the assessments as implementation systems mature especially as they can be carried over into other examinations and assessment processes. Countries should therefore adopt an incremental improvement standardisation strategy to strengthen the credibility of assessment results. Standardisation of assessment testing may focus initially on issues of administration and process fidelity, and then on assessment content, analysis and reporting to secure progressive improvements in the extent of standardisation of learning assessments.

33 Internal validity is defined as the extent of control over extraneous variables such that test design, instrumentation and procedures to eliminate possible sources of error. External validity refers to how generalizable the results are. Test reliability refers to consistency of measurement and the similarity of the results if the test is repeated on different occasions (Wagner, 2010).

Test design and development

Effective educational assessment requires test instrument design appropriate to the construct being measured. Lockheed (2008) confirms the requirement for assessment instruments designed to be stable over time periods T_1 to T_2, in at least six ways, all of which are met by international assessments and national assessments such as TIMSS and NAEP. She proposes the following in her research on tracking learning progress: testing the same cohorts (grade- or age-cohort) over time; measuring the same academic content, competencies or constructs over time (validity); using the same technical sampling procedures and methods (standardisation); using measurement instruments with constant levels of difficulty (implying a stable mix of items of differing cognitive difficulty levels); and using measurement instruments having the same reliability so that the tools and items for measurement in the jurisdictions do not exhibit bias. Linn (2005) supplements Lockheed's findings and includes equivalence through the empirical equation of tests using linked or common items and statistical methods to equate the tests.[34]

Lockheed, in her research on measuring learning progress over time confirms the need for overall item difficulty patterns should remain the same over time and should be confirmed using psychometric methods applied to the responses of the tested populations so that changes in test score reflect changes in learning over time. Test equation requires the use of anchor items which are common between years. Psychometric capacity to empirically equate the tests is also essential although both are in short supply in most developing countries.

[34] Tests should be capable of being equated across grades (vertically) in terms of their ability to determine the amount of learning that has taken place by a cohort when tracked or equated between years (horizontally) in terms of actual changes in learning given knowledge about the relative ease or difficulty of the test paper from year to year. The easiest way of comparing test score growth through different tests or forms of test is to include common or 'anchor' items in different forms of test administered to different grades or at different points in time. This enables the performance of a particular set of learner on those items to be assessed and used to scale the other test results of other learners with different performance levels. For example, if in one test the highest performers are from well-resourced schools and their performance in two years differs, then all other changes in the performance of other students will be interpreted or scaled to that difference from one year to the other. This equation of test results should be done to establish the expected and maximal change possible. Performance should then be scaled or pegged in terms of this maximum performance standard or change.

Lockheed (2008) also notes that unstable tests which are not comparable in terms of test design and difficulty or in terms of construct measurement over time result in noisy data as in the Jamaican annual assessment data. Psychometric methods such as Item Response Theory (IRT) are used in NAEP and other international assessments to improve the empirical calibration of item difficulty and test attributes (Bandeira de Mello et al, 2015). This ensures that test results have the same meaning from one year to another although IRT methods have themselves been criticised by some authors involved in educational assessment (Greaney & Kellaghan, 2008; Dunne, Long, Craig and Venter, 2012).

Item development

Research by Cresswell et al (2015) indicates the best practice in item development which involves many steps including: item generation, panelling (review by technical specialists in item development based on information on item performance in pilots), cognitive trialling (content and language testing among target population), field trialling (to test for cultural appropriateness and any bias of the remaining items in the field) and selection for the main study. This ensures rigour, buy-in and integrity of item performance among the target population, provided that it is informed and guided by experts in the field of assessment implementation, item development and how the item behaves in tests of the target population using psychometric techniques. To minimise bias in item performance during the field trialling, items should be subjected to Differential Item Functioning (DIF) analysis as in NAEP, PISA and TIMSS. Cresswell et al (2015) indicate that all aspects of item development including field trialling, item functioning analysis, adjustments and analysis are typically documented in advanced assessment programmes.

Technical sampling integrity, efficiency and design

The costs and benefits of an assessment determine the sampling approach to be used. Test burden is a real concern especially in assessment systems which are not optimally aligned or linked. Tests cannot be too long or too taxing on the learners or they will confound the results (Mullis, Martin, Foy & Arora, 2012; Foy et al, 2007; Ross et al, 2005). Matrix sampling techniques in TIMSS use booklets with different

and overlapping questions administered to a sample of learners to reduce burdensome tests and learner test fatigue. In SACMEQ, the test is taken by all participants and its length is restricted as a result.

Sample-based assessment tests, though cheaper than universal assessments, are limited as they cannot be used for fine-grained diagnostic analysis of learner strengths and weaknesses. However, sample-based assessment tests can allow test security to be limited to certain geographical areas or to a service provider as with NAEP (Bandeira de Mello et al, 2009). In most sample-based testing, the same age or grade may be identified for testing at critical points in a schooling system over time and a sample randomly drawn[35]. As resources are finite, there may be some alternation between grades over years. Some assessments like PISA focus on learning at a single age although considering the school participation profiles, learners may be spread over more than one grade at any one particular age.

Sampling may lead to some exclusions due to the requirements of the survey design. For example, since small schools are excluded in SACMEQ, countries which have a large number of such schools relative to larger schools, should subject such schools to deeper study to understand the learning and teaching dynamics in such schools. This may be done through purposive studies, qualitative analyses and oversampling of such schools in national assessments.

Sampling decisions need to be made with sufficiently broad based consultation and be justified in terms of the need for trends in performance, preferably within the context of an assessment plan which is well understood, not just within the bureaucracy but within the sector. If these samples are to be representative at the provincial level, the sample size must be adjusted and error measurements must be

[35] Random sampling is used to determine a national sample of learners, with the administered tests scored and weighted, and used to make generalizable conclusions about the academic performance of the general student population. To relate the results from the sample to those in the general population from which the sample is drawn, sampling weights are applied to the results generated by the sample-based administration of any test. Simple statistical methods using the achievement scores cannot be used without adjusting for the representation of the whole population drawing from the study sample estimates.

included in the reporting. Although the technical aspects of sampling are not examined in detail in this study, they clearly require attention as some of the earlier assessment programme samples were not consistent across years or relied on rudimentary rules of thumb for sampling (early systemic evaluation, for example).

Analysis of assessment data

Typically, items are scaled in learning assessments so that learner scores are transformed and standardised to a score distribution with an average of 500 and standard deviation of 100. Item difficulty between years must be constant. This is the case for standardised assessments such as SACMEQ, TIMSS and PIRLS. According to Bandeira de Mello et al (2009; 2015), NAEP score scales are created via IRT and scale score distributions are estimated for groups of students. When the score scales are created, parameters describing the item response characteristics are estimated. NAEP is not designed to report individual test scores but rather produces estimates of scale score distributions for groups of students. The resulting scale score distributions describing student performance are transformed to a NAEP scale and summary statistics of the scale scores are estimated. Statistical tests are used to make inferences about the comparisons of results for different groups of students or for different assessment years. NAEP scale score distributions are described via achievement levels or item mapping procedures (Bandeira de Mello et al, 2009)[36]. Based on responses in the sample-based assessment, achievement levels and levels of disaggregation of the reported test scores are determined for different groups of students in the country. Full completion of all test items by all learners in all domains is therefore not necessary for systemic monitoring. Scores may be derived using psychometric techniques applied to matrix-sampled test administration using multiple booklets. The communication of scores at country level requires careful

[36] NAEP analysis consists of student background information and test item responses for each student – but inferences made only on populations or groups of students, thus no individual student scores are produced. Multiple plausible values are drawn for each student using information from the observed cognitive data (or test items) along with responses from the student, teacher, and school survey questionnaires. These values are produced by first scaling responses to the cognitive test items using item response theory rather than classical test theory, regressing the achievement within a subject/subscale onto survey data, and then drawing plausible values from the posterior distribution incorporating parameters estimated from the first two steps (NCES, 2015; Bandeira de Mello et al, 2009)

attention especially in terms of the historical understanding and reliance on classical scoring of tests.

Use of assessment data: considerations in reporting and dissemination

Aggregated reporting on progress in learning is routinely used many industrialised countries (Rosenkvist, 2010) and ranking is routinely and explicitly avoided due to the unintended behavioural consequences of ranking and test-based school accountability. Within the OECD, selective reporting of test scores and reporting by industrialised countries for the purposes of accountability is usually at aggregated level, disaggregated by groups of students (gender, disability status or location, for example) or reported by locale with the intention of informing and directing remediation nearer the school. Official EU documents in Austria, Belgium (the French Community), Denmark, France (in the case of *évaluations-bilans)* and Ireland state clearly that national tests cannot be used to rank schools (Rosenkvist, 2010). This is presumably in order to avoid the worst effects of indiscriminate school ranking observed in Margaret Thatcher's Britain and in response to well acknowledged research evidence which shows that test scores are influenced by home, school and learning context, and previous learning.

In response to international assessments such as PISA and national development priorities, countries such as Brazil have led the way in improving their test-based accountability systems over a period of years. Brazil developed a school quality index in 2005, a decade and a half after major education reforms started, using universal assessment data from public schools as the basis for the index and various iterations of improvements and data quality checks.[37] South Africa has indicated the desire to develop a basket of indicators for sector performance as advocated by researchers

[37] This universal assessment used the same instruments administered for a multi-subject sample-based systemic monitoring survey called the SAEB (Sistema Nacional de Avaliação da Educação Básica) in language and mathematics every two years to all learners in grade 5 and 9. It provides the information on which to develop the Índice de Desenvolvimento da Educação Básica (IDEB), a school quality index that combines student performance (as measured by Prova Brasil) and takes into account exclusion and repetition rates (de Castro, 2012).

including Koretz (2002, 2003 and 2008) and Conley (2015) in reports on performance in the education system (DoE, 2003a; DBE, 2010a & DBE, 2015a).

Reporting of education system progress is usually tied to annual and electoral cycles. Reducing the turn-around times of systemic assessment studies can enable better and more relevant feedback, information and use of assessment data for policy and accountability purposes within such cycles. Long feedback times compromise the utility of learning assessments and policy relevance in terms of opportunities to affect instruction, policy and action.

Table 5 illustrates the time between fieldwork and reporting for some assessments in South Africa. It should be noted that in terms of turn-around time, sample-based ANA has been particularly rapid, due in part to the use of external capacity sourced for the purpose (SAB & T, 2013).

In OECD countries, a combination of universal and sample-based national learning assessments at the end of key stages of schooling is typically used (Rosenkvist, 2010). Reporting methods for accountability may include report cards at school and local level, with individual student report cards used at school level for local accountability purposes. To fill school level information and accountability gaps, South Africa's Department of Basic Education developed templates for ANA report cards to parents and school governing bodies (DBE, 2015a).

Table 5: Time taken to release assessment reports on South Africa's participation in various learning assessments

South African participation in national and international assessments	Fieldwork and data collection (month and/or year)	Report released in country (month and/or year)
ANA – universal 2012	Sep-12	Dec-12
ANA – V 2013	Sep-13	Dec-13
ANA – universal 2013	Sep-13	Dec-13
Systemic evaluation 2001	2001	2003
Systemic evaluation 2004	2004	2005
TIMSS 2002	Sept/ Oct 2002	Dec-03

TIMSS 2011 TIMSS 2015	Aug-11 Aug/ sept 2015 (Gr 9 data) and 2014 (Grade 5 data)	Dec-12 Nov-16
PIRLS 2006 PIRLS 2011	2005 2010	2007 2012
SACMEQ 2000 SACMEQ 2007 SACMEQ 2013	Sep-00 Sep-07 Sep-13	2005 2010 Jun-16 for preliminary results and August 2017 for final country report
US NAEP included for comparison	January to March	6 months from the day scoring starts

Sources: National Examinations and Assessment Directorate, Department of Basic Education (for SACMEQ and ANA), Pretoria. Human Sciences Research Council (for TIMSS). Centre for Educational Assessment, University of Pretoria (for PIRLS).

Large amounts of NAEP data and reports are available electronically on the web and dissemination of NAEP resources is done in online and print form for sub-national jurisdictions and public consumption. Brochures, reports, interactive results maps, and over 100 report card webpages, NAEP Data Explorer, NAEP State Comparisons, NAEP State and District Profiles, NAEP Questions Tool, released items, item maps and many pages of technical documentation on the NAEP assessments are disseminated over the web. The NAEP process, results and tools are explained separately for different groupings of stakeholders (educators, parents and students), with the educational implications clearly spelled out in terms of limitations of the data and results, possible utility of the results and the schedule of plans for the next cycle of assessments.[38] Such sophistication and variety in reporting is the result of decades of experimentation and communication in the NAEP and can be built into countries' long-range plans for assessment system research development and innovation. In addition, it must be clear what the different levels of reporting mean for the system in a standardised assessment.

Performance standards or achievement levels.

[38] www.nationsreportcard.gov

According to the National Centre for Education Statistics (NCES), NAEP performance standards are set every few cycles in a consultative process, with a panel of professionals making judgements on the three proficiency levels for each subject and test, with full psychometric and curricular information on the constructs measured by each item and in the tests.[39] The National Assessment Governing Board convenes a panel of experts to help set performance or achievement levels. The panel considers all items administered in the assessment and attempts to place them on the scale in terms of the cognitive requirements for performance, based on a series of technical inputs on the psychometric attributes of the items. For example, using information from psychometric item and person performance profiles, panellists are tasked to place Grade 4 items that requires students to read a word problem that requires the identification of a specific mathematical operation (such as addition with regrouping) to solve the problem and then give an answer. On a scale of 0 to 300, some panellists may assume that this is a very difficult problem and place it at 200. Through several iterations, negotiations and understanding of test item performance, all items are considered until there is agreement on the location on the scale of the responses for the item, and on the levels of knowledge are required In this case, technical information from the item performance and person proficiency information is combined with professional knowledge of the curriculum and assessment items for every subject assessment framework in the NAEP collection[40].

Federal systems may need to lay national benchmarks for proficiency using assessment tests – especially when national and sub-national assessment arrangements exist as is the case in South Africa. In a once-off exercise, national achievement levels were also used to test the definitions of "proficiency" at state level using NAEP and NAEP levels of proficiency were found to be more exacting than state measures (Bandeira de Mello et al, 2015). These research findings contribute to

[39] Personal communication on NAEP with Dr E. Sikali, National Centre for Education Statistics (NCES), November 2015

[40] Discussion with NCES staff on 7 October 2015 including Eunice Greer: reading test development and scoring; Elvira Hausken: test development oversight and maths; Lauren Harrell: Psychometric measurement; Emmanuel Sikali: Training and capacity building in data analysis; Sheila Thompson: National PIRLS coordinator; Bill Ward: Sampling and Data collection; Jamie Deaton: Questionnaire design; Dan McGrath: Reporting and dissemination; Dave Test: Special Education.

benchmarking learning proficiency and contrasted with reports of improving sub-national state performance in state-administered assessment tests reported by the US states themselves according to Darling-Hammond & Wentworth (2010).

Measuring test score growth

The methods of reporting test score changes from assessment exercises provides signals for how further test score improvement may be secured, as it can influence quality-seeking behaviour in the system positively or negatively. The province of Alberta embeds the test scores as one of the pillars of accountability within a multi-indicator framework for school performance, encouraging a more nuanced concept of school improvement (Alberta, 2013).

According to Hull (2007), reporting the grade-level students scoring at a particular level of performance in the assessment year is referred to as status growth reporting. This may lead to diversion of resources away from students at the lower and higher end of the performance spectrum to those on the margins of particular scoring bands, resulting in more inequity. Hull further notes that growth models, by contrast, emphasise the reporting of progress in academic achievement by individual learners over time and are more often used to stimulate quality-seeking behaviour. For example, reporting mathematics score improvement for an appropriately scaled grade-appropriate test may focus on the increase in score of a Student X by fifty points from 300 in Grade 5 to 350 in Grade 6.[41]

Typically, test score growth is measured in terms of the number of standard deviations represented by the change in test score means in a country's education system, whether from sample-based systemic learning assessments or from universal assessments. The standardisation of scores pegs the standard deviation of

[41] **Scale score:** A single numeric score that shows the overall performance on a standardized test. Typically, a raw score (number of questions answered correctly) is converted to a scale score according to the difficulty of the test and/or individual items (for example, the 200–800 scale used for the SAT.)
Vertical scale scores: Numeric scores on standardized tests that have been constructed so that the scale used for scoring is the same for two or more grade levels. Hence, a student's scale score gain over multiple years represents the student's level of academic growth over that period of time.

100 for standardised scores making score growth easy to compute over years[42]. Bruns, Evans & Luque (2012) indicate that the 52 score improvement in Brazil in PISA from 2000 to 2009 was of the order of 0.04 standard deviations per year although lower than most OECD participant countries.

In trying to approximate what this test growth means in terms of learning progress, Reddy, Prinsloo et al (2012) quote TIMSS estimates which show that within a 4-year testing cycle, a country could improve at best, by one grade level or year of learning, equivalent to 0.4 of a standard deviation for TIMSS. Around half a standard deviation is therefore taken as a reasonable estimate of the amount of learning which should take place in a schooling year. Hanushek & Woessman (2007) approximate possible improvements in education at 0.5 standard deviation per decade (or 0.05 standard deviations per year).

In evaluation research, there are limitations to representing test score growth in units of standard deviations (Vivalt, 2015). Despite assumptions that this is the case, this dispersion is not always constant across populations in the comparison of countries' scores in international assessments (Singh, 2015). Tests may exhibit floor effects, with a large proportion of students with very low scores; or the test may be too easy and may not adequately discriminate between high performers. Both have the same effect of exaggerating test core growth (calculated in terms of standard deviations). Compared with a country with a more homogenous group of test takers using the same test, test score growth appears higher in a country with a heterogeneous group of test takers due to the larger dispersion of test scores around the mean.

At learner level, significant effect sizes in isolated education programmes may increase test scores by the order of around 0.2 standard deviations over the life of the intervention. According to Mohohlwane (2016), for example, McEwan (2015) identified the interventions which had substantial effects, in units of standard

[42] **Standard deviations** (SD) indicate variation within a distribution the higher the variation in scores, the more spread around the average. Standardised assessments typically have a fixed SD of 100. Any improvements in the point scores can therefore be calculated in terms of SD per year. Brazil's 9 year improvement of 52 points is calculated as (52/100)/9 SD per year = 0.06 standard deviations per year.

deviations, on learner achievement. These were computer or instructional technology (0.15); teacher training (0.12); smaller classes with ability grouping (0.12); contract and volunteer teachers (0.10); student teachers or performance incentives (0.09); and incorporating instructional materials (0.08). Based on a systematic review of 18 studies of education interventions, Snilstveit, Stevenson, Phillips, Vojtkova, Gallagher, Schmidt, Jobse, Geelen, Pastorello & Eyers (2015) indicate that the largest and most consistent positive average effects on learning achievement were brought about by structured instructional improvement programmes consisting of content focused on deep understanding of a topic, instructional and learning support materials and intensive teacher training focused on the topic.

When analysing and reporting improvements, due consideration must be given to the target population of test takers and their profile, and the measures of test score improvement used must be appropriately chosen to ensure credibility.

Summary

This section of the chapter presented the evidence for the dimensions of Clarke's original and modified framework which was used to modify the tool for analysing and evaluating systemic learning programmes. Table 6: below summarises the differences between the dimensions and sub-dimensions in the original and modified frameworks for analysis with evidence for the modifications from the literature. The review of best practice and literature confirms the utility and support for systemic learning assessment in monitoring education and curriculum investments over time. Information provided by such systemic learning assessments, if provided in a timely fashion and in the correct formats underpinned by technical standards for the administration of assessments, can facilitate improved feedback and better monitoring of trends in a country's learning outcomes. In order to eliminate inequity, the unintended consequences of accountability measures need to also be monitored and minimised to improve the credibility of the results of monitoring learning progress. This can, in turn, enable the provision of improved quality of basic

education to the public, to learners and their parents and to education practitioners, researchers and policy makers.

Table 6: Evidence, for Clarke's original framework and modifications to the SABER rubric, which is used to analyse programmes intended to measure learning progress at the country level.

Dimension in the original and modified framework/ rubric	Sub-dimensions	Evidence for sub-dimension from the literature
Enabling context (original framework)	Social validation Institutional and policy context Sustainable funding Integrity of organisational structure, capacity and human resources	Enabling Context 1: Setting clear policies for NLSA/policy and institutional context (Ferrer, 2006; de Castro, 2012)
		Enabling Context 2: Having strong public engagement for NLSA/social validation (Hanushek and Raymond, 2005; Cartwright, 2013)
		Enabling Context 3: Having regular funding for NLSA/sustainable funding (Hoxby, 2002; Ferrer, 2006; Lockheed, 2008)
		Enabling Context 4: Having strong organisational structure for NLSA/organisational structure (Lockheed, 2008; Kellaghan and Greaney, 2008)
		Enabling Context 5: Having effective human resources for NLSA/human resources and capacity (Kellaghan and Greaney, 2008; Chakwera, Khembo & Sireci, 2004)
System alignment (original framework)	Extent of learning assessments alignment to curriculum Teachers (and other partner's) understanding of learning assessment	System Alignment 1: Aligning the NLSA with learning goals/teacher and stakeholder understanding and curriculum alignment (Rosenkvist, 2010; Ross et al, 2005)
		System Alignment 2: Providing teachers (and others in the sector) with opportunities to learn about the NLSA (Darling-Hammond and Wentworth, 2010. Black and Wiliam, 1998; Hoadley, 2012; Taylor, 2011; Carnoy Chisholm & Chilisa, 2012)
Technical assessment quality (original framework)	Inclusiveness of the assessment. Use of assessment data: considerations in reporting and dissemination; Review and evaluation of effects of assessment on education system; Technical documentation and knowledge management.	Assessment Quality 1: Ensuring the quality of the NLSA/inclusive assessment and presence of technical documentation (Joncas and Foy, 2011; Jerrim, 2013). Assessment Quality 2: Ensuring effective uses of the NLSA/use, reporting and dissemination (Cresswell et al, 2015).
Modified assessment quality dimension (modified)	Security and confidentiality concerns. Standardisation of learning assessments: reducing bias Item and test design and development. Analysis and use of assessment data; considerations in reporting and dissemination; Item development and piloting; Technical sampling integrity, efficiency and design; Use of performance/ achievement levels; and, measurement of score growth	Assessment Quality (AQ) modification 1: Ensuring security, confidentiality and standardisation of NLSA/systemic assessment tests (Lockheed, 2008; Koretz, 2013) AQ modification 2: Having well designed and developed tests (Conley, 2015; Cartwright, 2013; Linn, 2005; Greaney and Kellaghan, 2008) AQ modification 3: Ensuring effective and appropriate item development Cresswell et al, 2015) AQ modification 4: Ensuring sample integrity, efficiency and design (Ross et al, 2005; Lockheed, 2008) AQ modification 5: Ensuring technically valid analysis, reporting and dissemination of learner performance and achievement levels (Koretz, 2002, 2003, 2008; Hull, 2007; Bandeira de Mello et al, 2009; Vivalt, 2015)

2.10 Conclusion

This chapter has presented the evidence for best practice in measuring learning progress. There is remarkable agreement in the literature about the nature of the technical requirements for implementing learning assessments in general and especially those required for measuring progress in learning. Clarke's framework, and the modifications proposed in this study, supports a developmental approach to elaborating educational assessment systems. The framework is consistent with the findings in the literature and provides the opportunity to deepen the effectiveness of educational assessment systems and reform at the country level. The dimensions of the framework allow for the specification of different aspects of the enabling context and system alignment within the education system-schooling sector. The consensus in the literature of education reform, policy implementation and educational system change confirms that technical considerations are not sufficient conditions for successful implementation according to Allen, Elks, Outhred and Varly (2016); Best et al, 2013). Any change at system level of educational reform has to be done thoughtfully and patiently, with sufficient consideration of technical assessment understanding and instructional utility.

Examination of the literature and best practice supports the use and modifications to Clarke's framework (2012b) and the evaluation rubric associated with the framework developed by SABER (2014). The slight modifications to the SABER rubric emphasised the expansion of the system alignment sub-dimension criteria to include other officials in addition to teachers who have a deeper understanding of the different assessment functions, types, results, and links to other parts of the system like curriculum and teacher development. The assessment quality dimension was modified to emphasise the technical aspects of implementation arising from South Africa's recent implementation of various programmes include the standardisation of test administration and other processes, confidentiality and security specifications, test design, analysis and sampling, as well as the use of performance levels and methods to analyse and report test score growth. The modification of Clarke's framework and the SABER rubric carried out in this study is justified as it provides

practical detail for where governments in developing countries such as South Africa should invest in order to more meaningfully measure the progress of learning outcomes. For example, supporting test and item development using psychometric means for to ensure test difficulty stability and equivalence would improve the low ratings attached to test design in all examinations and assessments in the system, provided that these efforts are rigorously carried out. Better test design, item development, calibration and piloting, and comparability are crucial in sustaining future systemic learning assessment efforts in South Africa. The importance of technical support in the areas of weakness identified in the analysis in this study cannot be overstated. The modifications recommended in this study make more explicit some technical specifications which may have been implicit in the assessment quality dimension of Clarke's framework. They arise from the technical concerns of assessment arising from South Africa's experiences and they focus on the specific technical requirements of assessment quality interest to developing countries. As such, these modifications may be used to emphasise, or incorporated into the existing systemic alignment and assessment quality dimensions of the SABER rubric and Clarke's framework if required, to emphasise the technical aspects of implementation.

The next chapter presents an analysis and evaluation of the systemic learning programmes in South Africa between 1994 and 2016, using the dimensions of Clarke's modified framework and the evaluation matrix arising from the modified SABER rubric as the basis for analysis. Systemic learning assessment programmes are analysed comprehensively and evaluated in Chapter 3 with the findings presented thereafter.

Chapter 3 Evaluation and analysis of systemic learning assessments in South Africa.

This chapter presents the analysis of all of South Africa's systemic learning assessment programmes purporting to be systemic learning assessments for monitoring learning progress at the country level. The analysis was carried out using an evaluation matrix in Table 8 which is based on amendments to the SABER rubric (SABER, 2014). The detailed analytical framework and rubric helped to identify areas of weakness and strength in past and current systemic assessment programmes, based on the three dimensions of enabling context, system alignment and assessment quality (Clarke, 2012b). The resulting analysis thus allows specific and practical recommendations to be made on where and how to strengthen these programmes so as to effectively track trends in learning outcomes in South Africa. Judgements in the evaluation were made using documented information, literature in the public domain, along with departmental documentation which was unpublished. Information was also collected from individual key informants in areas where institutional memory was required, information was difficult to gain access to, or where documentation on the issue would gave breached security protocols.

3.1 Analysis of systemic learning assessments in South Africa: ratings and evaluation method

The modifications to the SABER evaluation matrix arising from this study are presented in bold font in Table 8 below which spans multiple pages, along with the rating of the level of development of each systemic assessment programme in South Africa. In the tables in this chapter, *'Enabling context'* refers to the overall policy and resourcing framework within which classroom assessment activity takes place in a country or system. *'System alignment'* refers to the degree to which the assessments are coherent with other education system components. *'Assessment quality'* refers to the quality of the underlying technical components and processes which make up the assessment (Clarke, 2012b). Table 7 below is an extract of Table 6 in the previous

chapter and illustrates the dimensions and sub-dimensions required for effective student assessment programmes.

Table 7: Dimensions and sub-dimensions of systemic learning assessment programmes with modifications arising out of the findings of this study, modified from evaluations rubric developed by SABER (2014) and based on Clarke (2012a; 2012b)[43]

Dimensions	Sub-dimensions
Enabling Context	Enabling Context 1: Setting clear policies for NLSA / policy & institutional context
	Enabling Context 2: Having strong public engagement for NLSA/ social validation
	Enabling Context 3: Having regular funding for NLSA/ sustainable funding
	Enabling Context 4: Having strong organisational structure for NLSA/ organisational structure
	Enabling Context 5: Having effective human resources for NLSA/ human resources and capacity
System Alignment	System Alignment 1: Aligning the NLSA with learning goals/ teacher and stakeholder understanding and curriculum alignment
	System Alignment 2: Providing teachers (and others in the education enterprise) with opportunities to learn about the NLSA
Assessment Quality	Assessment Quality 1: Ensuring the quality of the NLSA/ inclusive assessment and presence of technical documentation
	Assessment Quality 2: Ensuring effective uses of the NLSA/ use, reporting and dissemination
Assessment Quality Modification	Assessment Quality Modification 1: Ensuring security, confidentiality and standardisation of systemic assessment tests
	Assessment Quality Modification 2: Having well designed and developed tests
	Assessment Quality Modification 3: Ensuring effective and appropriate item development
	Assessment Quality Modification 4: Ensuring sample integrity, efficiency and design
	Assessment Quality Modification 5: ensuring technically valid analysis, reporting and dissemination of learner performance/achievement

Ratings in Table 8 below were allocated in the analysis using an evaluation of the assessment programme based on the evidence from the literature, policy and technical documentation, and occasional personal interviews on the particular sub-

[43] Systemic learning assessments are called National or System-Level Large Scale Assessments (NLSA) in Clarke's original framework and the SABER rubric (Clarke, 2012b; SABER, 2014) and they are included in Table 8.

dimension being analysed. Ratings were carried out using a colour coded system (with red denoting latency or absence of the attribute or sub-dimension; amber/orange denoting emerging status; light green denoting established status; and, dark green denoting an advanced state of development). Equal weighting is given to each of the three dimensions of enabling context, system alignment and quality which were examined in the evaluation matrix. Averaging the ratings across the sub-dimensions and dimensions respectively yielded a mix of colours related to mean ratings calculated.

The numerical value of each rating calculated or allocated is included for clarity. Bold format in the evaluation matrix in Table 8 indicates additions to the original SABER rubric and the square brackets with text struck out indicates text from the original matrix which has been deleted and replaced by bold formatted insertions. ANA referred to in the table is the sample-based version of the Annual National Assessment, along with the Trends in Mathematics and Science Study (TIMSS), the Monitoring Learning Achievement (MLA) study, the Southern and Eastern African Consortium for Monitoring Education Quality (SACMEQ) programme, the Systemic Evaluation (SE) programme, the National School Effectiveness Study (NSES), the Progress in International Reading Literacy Study (PIRLS) and the National Senior Certificate (NSC).

Table 8: below includes the detailed evaluation matrix populated with ratings from a modified SABER rubric. A latent score was used where there was an absence of, or deviation from, the attribute being rated. For example, once-off assessments such as the MLA, NSES are rated at 1 for public engagement, social validation and generalised teacher and stakeholder understanding as these two programmes were carried out only once as research or benchmarking exercises. Other levels of development used in the ratings include emerging rated at 2, established rated at 3, and advanced rated at 4. The overall average rating for each of the three different dimensions of systemic learning assessments are presented showing the summarised level of development for the different programmes, which are retained as in the SABER rubric as National Large Scale Assessments (NLSA) in

Table 9 below.

Table 8: Evaluation matrix with rating rubric and criteria for evaluating the level of development of systemic assessment programmes in South Africa - modified from Clarke (2012b Annexure A) and SABER (2014)[44]

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
Enabling Context 1: Setting clear policies for NLSA/policy and institutional context				3.3[45]	3.0	1.0	3.0	1.8	3.0	1.0	2.3
No National Large-Scale assessment (NLSA) exercise has taken place	The NLSA has been operating on an irregular basis.	The NLSA is a stable program that has been operating regularly.	This option does not apply to this dimension.	3	3	1	3	1	3	1	2
There is no NLSA policy document	There is an informal or draft policy document that authorises the NLSA.	There is a formal policy document that authorises the NLSA.	This option does not apply to this dimension.	3	3	1	3	3	3	1	3
The policy document is not available outside the	The policy document is not available to the public.	The policy document is available to the public.	This option does not apply to this dimension.	3	3	1	3	2	3	1	2

[44] Bold indicates additions to the original rubric and framework based on the study findings. Square brackets indicate the original text deleted from the rubric and replaced with bold text additions forming the evaluation matrix for this study. ANA referred to in the table is the sample-based version of the Annual National Assessment, TIMSS is the Trends in Mathematics and Science Study, MLA is Monitoring Learning Achievement study, SACMEQ is the Southern and Eastern African Consortium for Monitoring Education Quality, SE is Systemic Evaluation, NSES is the National School Effectiveness Study, PIRLS is the Progress in International Reading Literacy Study and NSC is the National Senior Certificate.

[45] Average ratings calculated per sub-dimension are included, for enabling context, system alignment, and assessment quality, in colour and figures.

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
implementation unit [Original: ~~This option does not apply~~]											
There is no plan for NLSA activity.	**There is some understanding of the need for an NLSA [~~Original: This option does not apply~~]**	There is a general understanding that the NLSA will take place.	There is a written NLSA plan for the coming years.	4	3	1	3	1	3	1	2
Enabling Context 2: Having strong public engagement for NLSA/social validation				3.0	3.0	1.0	3.0	1.0	3.0	1.0	2.0
Stakeholder groups strongly oppose the NLSA or are indifferent to it.	Some stakeholder groups oppose the NLSA.	Most stakeholders groups **are aware of and** support the NLSA.	All stakeholder groups support the NLSA.	3	3	1	3	1	3	1	2
Enabling Context 3: Having regular funding for NLSA/sustainable funding				3.5	3.5	1.0	3.0	1.0	3.5	1.0	2.5
There is no regular funding allocated to the NLSA.	There is irregular funding allocated.	There is regular funding allocated to the NLSA.	This option does not apply to this dimension.	3	3	1	3	1	3	1	3

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
Funding implies sustainability and *ad hoc* surveys are rated as latent. ~~[Original: This option does not apply.]~~	Funding covers some core NLSA activities: design, administration, analysis and reporting, but it does not cover research and development.	Funding covers all core NLSA activities	Funding covers research and development activities in addition to core activities.	4	4	1	3	1	4	1	2
Enabling Context 4: Having strong organisational structure for NLSA/organisational structure				3.0	3.3	1.0	3.3	1.0	3.3	1.0	3.3
There is no NLSA office, *ad hoc* unit or team.	The NLSA office is a temporary agency or group of people.	The NLSA office is a permanent agency, institution or unit.	This option does not apply to this dimension.	3	3	1	3	1	3	1	3
There is no on-going technical interaction of the programme	Political[46] considerations regularly hamper	Political considerations sometimes **influences** ~~[hamper]~~ technical considerations **in the**	Political **considerations never hamper** technical	3	4	1	4	1	3	1	3

[46] 'Political' in this case relates to interactions and conduct of government administration and public affairs and not to party political considerations. The high ratings of 4 is given to assessments which were carried out independently on behalf of the Department of Basic Education, with minimises the risk of exposure to influence.

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
with policy-makers ~~[Original: This option does not apply]~~	technical considerations.	**country assessment.**	considerations. **Independent service providers used in the country assessment**								
There is no on-going technical interaction between programme activities and education administration structures. ~~[Original: This option does not apply to this dimension].~~	The NLSA office is not accountable to a clearly recognized body.	The NLSA office is **administratively and technically** accountable to a clearly recognized body.	This option does not apply to this dimension.	3	3	1	3	1	3	1	3
Enabling Context 5: Having effective human resources and capacity for NLSA				3.0	3.0	1.0	2.5	1.0	2.0	1.0	2.0
There is no staff allocated for running an NLSA.	The NLSA office is inadequately staffed to effectively carry out the assessment.	The NLSA office is adequately staffed to carry out the NLSA effectively, with minimal issues.	The NLSA office is adequately staffed to carry out the NLSA effectively, with no issues.	3	3	1	2.5	1	3	1	2

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
The coordinating agency/country does not offer opportunities that prepare individuals for work on NLSA.	**The coordinating agency/country offers few opportunities in limited sites to prepare individuals for work on the NLSA** [~~Original: This option does not apply~~]	The coordinating agency/country offers some opportunities to prepare individuals for work on the NLSA.	The coordinating agency/country offers a wide range of opportunities that prepare individuals for work on NLSA.	3	3	1	2.5	1	1[47]	1	2
System Alignment 1: Aligning the NLSA with learning goals/ teacher and stakeholder understanding and curriculum alignment				3.3	2.7	1.0	2.3	1.3	2.3	1.3	2.3
There is no evidence that	**There is an attempt to align**	The NLSA measures performance against	This option does not apply	3	2	1 [48]	2	2	3	2	3

[47] Unlike TIMSS, PIRLS implementation service providers in the years up till 2015 provided limited opportunity for preparing officials and personnel outside the providers, unlike TIMSS which ensures that a select group of researchers is trained in addition to the service provider after each cycle. Access to information on country-specific African language data, instruments and performance information at the individual and item level in pre-PIRLS 2011 implementation, outside of the technical report, was difficult to source from the service provider in relation to the African languages (DBE officials' discussion with the head of the Centre for Education Assessment at the time and the Dean of Education Prof Eloff and Professor Duncan, DVC Academic affairs at the University of Pretoria on 17 November 2015 was unsuccessful in obtaining this information). The challenge of information sharing may have improved with a change in personnel at the Centre, however, African language data, tools and item level data for further analysis.

[48] No documentation on the constructs or learning standards used in the MLA instruments could be found in the public domain apart from the country report although the UNESCO report indicates some items and the responses to these items.

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
the NLSA is curriculum aligned [~~Original: It is not clear if the NLSA is based on curriculum or learning standards.~~]	**the assessment framework to curriculum standards.** [~~Original: This option does not apply~~]	curriculum or learning standards, **and includes empirical evidence of such alignment.**	to this dimension.								
There is no evidence of broad technical consultation on the constructs measured in the NLSA. [~~This option does not apply~~]	**What the NLSA measures is frequently questioned by stakeholder groups** [~~This option does not apply~~]	What the NLSA measures **is accepted by most officials and stakeholders and** is questioned by some stakeholder groups.	What the NLSA measures is largely accepted by stakeholder groups **and education officials at different levels**.	4	3	1	3	1	1	1	2
There are **few or** no mechanisms in place to ensure that assessments	**There has been more than one ad hoc reviews of the NLSA to ensure that it**	There are regular internal reviews of the NLSA to ensure that it measures what it is intended to measure.	This option does not apply to this dimension.	3	3[49]	1	2	1	3	1	2

[49] All assessments generally have review processes associated with the constructs being measured at country level or with TIMSS and PIRLS, according to IEA protocols.

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
accurately measures what it is supposed to measure.	**measures what it is intended to measure.** ~~[Original: There are ad hoc reviews to ensure valid measurement]~~										
System Alignment 2: Providing teachers (and others in the system) with opportunities to learn about the NLSA				3.0	2.0	1.0	3.0	1.0	2.0	1.0	3.0
There are no courses or workshops on the NLSA implementation in country.	There are occasional courses or workshops on the NLSA w**ith limited participation at country level.**	There are some courses or workshops on the NLSA offered on a regular basis **to teachers, education practitioners, planners and researchers**.	There are widely available high quality courses or workshops on the NLSA offered on a regular basis.	3[50]	2[51]	1	3	1	2	1	3

[50] The focus on education practitioner capacity building is evident in the ratings for the national assessments, while the budget constraints associated with international assessments may militate against such activities on a wide scale.

[51] PIRLS and TIMSS to more people involved in education in-country and to link with quality measurement in the education system. TIMSS is better-known as efforts at advocacy have been more visible to education researchers than PIRLS and the data is more accessible from the service provider. For example, TIMSS 2015 was shared at a Basic Education Sector Lekgotla 23 to 25 January 2017 as a result (according to the Department of Basic Education, Annual Report for 2016/17).

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
Assessment Quality 1: Ensuring the quality of the NLSA/ inclusive assessment and presence of technical documentation				3.3	3.0	1.0	2.7	1.0	2.3	2.3	2.0
All *ad hoc* surveys are rated as latent.	**No options are offered and documented regarding inclusion of all groups of students in the NLSA.** ~~[Original: This option does not apply to this dimension.]~~	At least one option is offered to include all groups of students in the NLSA.	Different options are offered to include all groups of students in the NLSA.	4	3	1	3	1	2	1	2
There is no mechanism to ensure NLSA quality	There are no documented **empirically based documented** mechanisms in place to ensure NLSA quality.	There are some **empirically based** mechanisms in place to ensure the quality of the NLSA	There are a variety of empirically **based** mechanisms in place to ensure the quality of the NLSA.	3	4	1	3	1	3	3	2

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
There is no technical report or documentation about the quality of the NLSA.	There is some documentation about the technical aspects of the NLSA, but it is not in a formal report format.	There is a comprehensive technical report and documentation but with restricted circulation in each cycle.	There is a comprehensive, high quality technical report available to the general public **on country level processes.**	3	2[52]	1	2	1	3	3	2
Assessment Quality 2: Ensuring effective uses of the NLSA/ use, reporting and dissemination				3.0	2.7	1.0	2.7	1.7	2.3	2.3	2.3
NLSA results are not disseminated **widely at sub-national level.**	NLSA results are poorly disseminated - **some stakeholders are not aware of results, especially at school level.**	NLSA results are disseminated in an effective way.	**NLSA results are used for deep/ value-added education research and analysis.** ~~[Original: This option does not apply to this dimension.]~~	3[53]	2	1	2	1	2	2[54]	2

[52] There was no technical report for the TIMSS 2011 survey at country level apart from the summary report of performance, hence the lower rating.

[53] The NSC has been the main indicator of school level outcomes in the last two decades but it requires supplementation of other indicators of school performance.

[54] This rating for the NSES dissemination is for appropriate dissemination of the results since it has been argued elsewhere in this report that the NSES was primarily designed as a research study and not a monitoring programme.

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
NLSA awareness is limited at country level. ~~[Original: NLSA information is not used, or it is used in ways inconsistent with the purposes or the technical characteristics of the assessment]~~	**NLSA information is not used, or it is used in ways inconsistent with the purposes or the technical characteristics of the assessment** ~~[Original: This option does not apply to this dimension.].~~	NLSA results are used by some stakeholder groups in a way that is consistent with the purposes and technical characteristics of the assessment.	NLSA information is used by all stakeholder groups in a way that is consistent with the purposes and technical characteristics of the assessment.	3	3	1	3	3	3	3	2
There is no evidence that the consequences of the NLSA are important to monitor for	**There are no mechanisms in place to monitor the consequences of the NLSA in the education**	There are some mechanisms in place to monitor the consequences of the NLSA in the education system.	There are a variety of mechanisms in place to monitor the consequences of the NLSA in the	3	2[55]	1	3	1	2	1	2

[55] With TIMSS and the NSC, there have been some attempts to gather information on the consequences of the collection on processes in the education system. The effects on instruction, school managers and other officials, as well as teachers' use of the assessment results in the system, remains largely unknown outside of a few case studies developed as postgraduate research theses.

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
planning. ~~[Original:There are no mechanisms to monitor the consequences of the NLSA in the education system.]~~	**system.** ~~[Original: This option does not apply to this dimension.]~~		education system.								
Assessment Quality Modification 1: Ensuring security, confidentiality and standardisation of NLSA/systemic assessment tests				3.0	4.0	1.0	2.0	2.0	4.0	3.0	2.0
There are no mechanisms for ensuring standardised test administration procedures, test security and/or confidentiality.	**There are some mechanisms for ensuring standardised test administration procedures, test security and/or confidentiality.**	**There are effective mechanisms for ensuring and checking deviation from standardised test administration procedures, and security/confidentiality protocols**	**There are well evaluated mechanisms for ensuring and checking standardised test administration procedures, test security and confidentiality**	3	4	1	2	2	4	3	2

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
Assessment Quality Modification 2: Having well designed and developed tests				2.0	3.0	1.0	2.0	2.0	3.0	2.0	1.5
Checks for bias are not applied or reported on in assessments in the system.	**Checks for bias are sometimes applied and reported on in the NLSA.**	**Checks for bias are always applied and reported on in detail in formal technical documentation. Sources of bias in the sampling of learners, schools and other participants are routinely quantified and reported on.**	**Fidelity checks for standardised assessment and checks for bias are used to develop improvement plans to increase quality assurance and oversight of the different aspects of assessment implementation.**	2.0	3[56]	1	2	2	3[57]	2	2
Tests are not comparable over different years	**Tests development includes some generally accepted standards and**	**Test development includes all generally accepted standards for test development and design. There is some evidence and**	**Tests developed are stable and involve the measurement of the same content,**	2	3	1	2	2	3	2	1

[56] TIMSS has bias-checking built into the standard procedures of the IEA which co-ordinates this international assessment. However, the lack of technical documentation on country-level activities, for 2011 in the public domain, makes it difficult to confirm this.

[57] PIRLS has standard bias checking built into the standard procedures of the IEA which coordinates this international assessment. However, little evidence of country-specific bias checks is available especially between the tests in different languages in PIRLS 2006 documentation.

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
	processes for test development, but little evidence or technical documentation to confirm test reliability, and comparability with time. There are some mechanisms to ensure that tests used are comparable from year to year, and cater for the full spectrum of learner abilities (from low to high performers)	**documentation to confirm that tests are reliable and unbiased and fit for measuring learning progress. There are effective empirical mechanisms to ensure that tests used are comparable from year to year, and cater for the full spectrum of learner abilities (from low to high performers).**	**standardised sampling methods, levels of difficulty over time. Psychometric and statistical evidence confirms that tests are reliable and unbiased and fit for measuring learning progress over time. Tests used are comparable from year to year, and cater for the full spectrum of learner abilities (from low to high performers)**								

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
Assessment Quality Modification 3: Ensuring effective and appropriate item development				2.0	3.0	1.0	2.0	2.0	3.0	2.0	2.0
No information on item development exists.	**Some aspects of the item development cycle are used. Summary descriptive documentation of item development processes exists.**	**The full item development cycle (of item generation, panelling, cognitive trialling, field trialling, and selection processes) is used, with iterations where appropriate. Technical documentation exists for all stages of item development in each cycle.**	**Item development innovations are linked to other assessment reform initiatives in the system including the generation of items banks. Technical documentation exists for all item development activities and decisions made. Mechanisms exist to evaluate and review these decisions from a technical and a governance viewpoint.**	2	3	1	2	2	3	2	2

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
Assessment Quality Modification 4: Ensuring sample integrity, efficiency and design				1.0	3.0	1.0	2.0	3.0	2.0	3.0	2.0
Sampling is not appropriate – this option applies to summative examinations for selection or universal assessment programme.	**Samples drawn with limited documentation on technical and cost considerations and decisions.**	**Clear mechanisms exist to ensure that sampling is done with sufficient technical understanding, and documentation elaborating on the sampling is produced for each cycle.**	**A high-quality technical report is available freely on the sampling, technical and cost considerations of the systemic assessment programmes. Mechanisms exist to translate the findings into strategies for improving sampling and analyses.**	1[58]	3	1	2	3	2	3	2

[58] The NSC examination is universal at Grade 12 and there is no need to sample, hence the latent rating.

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
Assessment Quality Modification 5: ensuring technically valid analysis, reporting and dissemination of learner performance and achievement				2.8	3.5	1.0	2.0	2.0	3.0	3.0	2.0
Basic classical scores growth methods are used to compute performance.	**Descriptive methods are sometimes used to determine and report on score growth.**	**Empirically defensible methods are generally used to determine and report on test score growth from year to year. Reports on statistical and psychometric checks are documented.**	**Empirically defensible methods are always used to determine and report on test score growth, and used to generate adjustments to the measurement of the progress in learning from year to year.**	3	4	1	2	2	4	4	2
Limited or absence of mechanisms to manage reporting and communication of the findings.	**Mechanisms exist to ensure reasonable reporting times for all aspects of the assessment.**	**Mechanisms exist to ensure improvements in response time for reporting of results to all stakeholders. Plans exist to improve efficiency and turnaround times and**	**Evaluations of administration and assessment processes yield continuous improvement strategies for test**	3	3	1	2	2	3	3	2

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
		costs without sacrificing test and result credibility.	**administration, analysis and reporting. Clear mechanisms exist for communicating and substantively linking these strategies with various internal and external stakeholders in the system.**								
One reporting format was used for country level reporting.	**Well-evaluated reporting formats are generated for use in communication to some groupings of stakeholders.**	**A variety of well-evaluated reporting formats are used for advocacy and communication of plans, activities and results for different levels and all groupings of stakeholders. Reports are packaged for a comprehensive set**	**Ongoing evaluation of stakeholder knowledge and perceptions of assessment system are used to generate dedicated strategies to improve future**	3	3	1	2	2	2	2	2

Level of development and rating				Systemic assessment programmes in South Africa: evaluation							
LATENT (Rating=1)	**EMERGING (Rating=2)**	**ESTABLISHED (Rating=3)**	**ADVANCED (Rating=4)**	**NSC**	**TIMSS**	**MLA**	**SACMEQ**	**SE**	**PIRLS**	**NSES**	**Sample-based ANA**
		of stakeholder groupings within and outside the system.	**communication and reports.**								
Learner performance reported using basic percentages or mean scores in summarised form.	**Some mechanisms exist to report on what learners know with some empirical / statistical rigour with appropriate disaggregation.**	**Mechanisms exist to ensure that empirically valid methods are used to compute and report on learner performance levels in relation to what they can do between years. Mechanisms exist each cycle for the use of achievement levels in reporting test results.**	**Mechanisms exist for evaluating, reviewing and generating further analysis to improve what is known about the relationship between learner achievement / test performance in relation to the intended curriculum.**	2	4	1	2	2	3[59]	3	2

[59] No documentation on all mechanisms available in PIRLS especially in relation to learner achievement in different languages in PIRLS 2006.

Table 9: Summary matrix with ratings for the level of development of systemic assessment programmes in South Africa. Modified version of the SABER (2014) rubric.

Dimensions of systemic learning assessments (also called National large Scale Assessments).	NSC	TIMSS	MLA	SACMEQ	SE	PIRLS	NSES	Sample-based ANA
	Examination for certification	Large scale international assessment	Research survey	Large-scale regional assessment	Survey	Large-scale international assessment	Research study	Systemic large-scale assessment at country level
Enabling context average level of development	3.2	3.2	1.0	3.0	1.2	3.0	1.0	2.4
System Alignment average level of development	3.2	2.4	1.0	2.6	1.2	2.2	1.2	2.7
Assessment Quality average of level of development (original framework and rubric)	3.2	2.7	1.0	2.7	1.4	2.2	2.3	2.0
Assessment Quality Modification average of level of development modified framework and rubric)	2.2	3.3	1.0	2.0	2.2	3.0	2.6	1.9
Average level of development for all dimensions of each assessment programme using the modifications in the study.	2.9	2.9	1.0	2.7	1.5	2.6	1.8	2.2

ANA referred to in the table is the sample-based version of the Annual National Assessment, TIMSS is the Trends in Mathematics and Science Study, MLA is Monitoring Learning Achievement study, SACMEQ is the Southern and Eastern African Consortium for Monitoring Education Quality, SE is Systemic Evaluation, NSES is the National School Effectiveness Study, PIRLS is the Progress in International Reading Literacy Study and NSC is the National Senior Certificate.

3.2 Analysis of systemic learning assessments in South Africa: discussion

This section discusses the analysis of systemic learning assessment programmes in South Africa and summarises the findings from this analysis and the literature. Programmes examined include the sample-based version of the Annual National Assessment (ANA), the Trends in Mathematics and Science Study (TIMSS), the Monitoring Learning Achievement (MLA) study, the Southern and Eastern African Consortium for Monitoring Education Quality (SACMEQ) programme, the Systemic Evaluation (SE) programme, the National School Effectiveness Study (NSES), the Progress in International Reading Literacy Study (PIRLS) and the National Senior Certificate (NSC) examination[60]. PIRLS is taken to include pre-PIRLS assessment as well, which is now called PIRLS Literacy.

Creating an enabling context for systemic learning assessment programmes is a basic requirement for implementation support. This requires sufficient institutional and organisational stability, policies and plans, political support, public and stakeholder engagement and validation, resourcing, accountability, and staffing of the programmes (Clarke, 2012b; SABER, 2014). The evidence for these sub-dimensions of enabling context is presented in the literature review in the previous chapter, and summarised in Table 6. The analysis indicates that technical and social validation and accountability is an area where systemic learning assessments are a weakness at the country level. The ANA 2015 crisis in South Africa arose from insufficient social validation (SADTU, 2014) as well as more fundamental technical shortcomings in the use and comparability of the results over time. The emerging to established rating of sample-based ANA reflects some of these concerns in the literature (DBE, 2011a; SAB&T, 2013). Some of these social and technical validation weaknesses are explored in more detail in the modified framework used in the analysis in Table 8. The research studies examined in this chapter are rated as latent due to their sustainability and limited cycles. NSES (Taylor, N., 2011), SE (DoE, 2003b, 2003c, 2005), and MLA (DoE, 1999), are examples of one-off or discontinued assessments.

[60] ANA (DBE, 2010a, 2011a, 2012b, 2013, 2015a), TIMSS (Reddy, 2006 & Reddy et al, 2012), MLA (DoE, 1999), SACMEQ (Moloi & Chetty, 2010; Ross et al, 2005), Systemic Evaluation (DoE, 2003b, 2003c, 2005), NSES (Taylor, N., 2011), PIRLS (Howie & van Staden, 2012; Howie et al, 2008) and NSC (DBE, 2014b, 2015c).

International studies which are rated as established such as the SACMEQ, PIRLS and TIMSS programmes indicate the substantial administrative and political support, resourcing and organisational support they enjoy within the National Department of Basic Education (DBE, 2017). Administrative burden and capacity were the main reason cited in opposition to the ANA processes which were halted in 2015 according to a departmental briefing of Districts on the new assessment reform processes by Dr M Chetty in 2017[61].

System alignment means that there are strong linkage between the assessment and learning goals, that the programme ensures that what is measured is valid and credible, that the results are accepted, and that quality assurance is sustained and that teachers are developed and understand the assessment. The recommendations of this study include modifying the second sub-dimension of system alignment to include ensuring the understanding of other education personnel, in addition to improving teachers' understanding of assessment system components, linkages and utility in assessment reform and educational improvement. There is compelling evidence to show a general lack of understanding in the education system, of the different functions of universal and verification ANA and the links between different types of assessment in education reform and development (Cartwright, 2013; DBE, 2015a).

As with the enabling context and assessment quality dimensions, once-off exercises such as Systemic Evaluation (SE), the Monitoring Learning Achievement (MLA), and National School Effectiveness Survey (NSES) are rated at or near latent as due to the absence of attributes such as the awareness, understanding and stakeholder and public engagement. Teacher and staff knowledge about the international surveys such as TIMSS and PIRLS is limited, especially at school level and this has been acknowledged in the historical literature (Taylor and Vinjevold, 1999) and in the current sector plan and documentation (DBE, 2015a). The established rating of the NSC (DBE, 2014b, 2015c) indicates that is widely understood at the country level and it is highly curriculum-aligned as it is used for certification purposes (Poliah, 2014).

[61] Dr M Chetty. Briefing of District Managers in the Department of Basic Education on 30 June 2017.

Many of the question setting panels, markers and moderators in the examination are drawn from the teaching corps, making it well known in the schooling system. The Annual National Assessment (ANA) is also curriculum aligned as it used panels of teaching professionals to set questions and enjoyed high levels of support especially after it was announced by the President of the Republic in 2011 (Zuma, 2011). The SACMEQ administration in-country also involves officials at district and circuit level (DBE, 2010a, 2011a, 2012b, 2013, 2015a), so there is some exposure of officials to these programmes.

Although there is frequently a trade-off between coverage of country-specific curricula and comparability in international assessments (Schiefelbein and McGinn, 2008), attempts have been made to accommodate country-specific questions in surveys such as TIMSS and PIRLS TIMSS (Reddy, 2006; Howie et al, 2008; Howie & van Staden, 2012; Reddy et al, 2012). The match is not always satisfactory as international assessment tests are not specifically aligned to curriculum standards at the country level (Lockheed, 2008). South Africa's review of TIMSS science items showed, for instance, that only a fifth of the items in TIMSS 2011 matched the National science curriculum of grade 7 students while half of the items matched the Grade 8 science curriculum (Howie & Hughes, 2000) although this has been reported to have improved to over 80% in many areas of the curriculum in the TIMSS 2015 exercise (Reddy et al, 2016). The regional SACMEQ assessment involved a documented country curriculum alignment process as part of implementation prior to 2005 (Ross, Saito, Dolata, Ikeda, Zuze, Murimba, & Griffin, 2005). There is, however, no official documentary evidence to confirm that this process was implemented again in subsequent SACMEQ cycles. However, verbal assurances from South Africa's SACMEQ National Research Coordinator confirm that this alignment exists for SACMEQ tests from 2013[62]. In terms of training programmes and workshops, these have been carried out using SACMEQ data by different academic institutions and research agencies as it provides opportunities for rich cross-country

[62] Personal communication with Dr M Chetty, Acting Director, National Assessment, Department of Basic Education, 2016.

comparison of learning outcomes according to the head of the SACMEQ Coordinating Centre[63].

Assessment quality requires that the assessment programme is inclusive, quality assured, and rigorous in terms of the technical analysis, with clear technical documentation of the implementation, reporting and the use of results (Clarke, 2012b). In addition, Clarke suggests that the use of results should be appropriate and technically defensible, that assessment results should be appropriately disseminated, and the consequences of the assessment monitored. The original assessment quality dimension looked at aspect of the use, utility and considerations in the reporting, application and dissemination of the assessments as well as knowledge management and the modified framework including technical elements of assessment. The modified framework emphasizes more technical issues and confidentiality concerns, the standardisation of learning assessments and the reduction of bias in results, as well as technical issues of item, test and sampling design, as well as reporting of test score growth and performance levels. It is acknowledged that the technical sub-dimensions and criteria added may have been implicit in the original framework (SABER, 2014), however, they are specified and emphasised because they relate to specific aspects of the South African experience of technical implementation of assessments, and they are included in the technical sub-dimensions in the rubric since they are relevant to developing countries in particular, and they may be incorporated into the relevant assessment quality sub-dimensions in the original framework.

The absence of technical documentation in SACMEQ for cycles after the second implementation of SACMEQ, outside of the regional country reporting formats, and the absence of a country-specific overall technical report for 2011 TIMSS are noteworthy aspects of assessment quality requiring remediation. Regional and international assessment programmes at country level especially TIMSS, PIRLS and SACMEQ should have a more comprehensive technical report with country-level processes and decisions documented. The evaluation and rating was focused on the

[63] Discussion with Ms T Masalila, Head of the SACMEQ Coordinating Centre, Botswana, 20 October 2016.

presence of comprehensive documentation at the country level available to help improve the understanding of the technical aspects of administration with a view to enhancing the understanding of decision- and policy-support in education reform efforts at the country level. By contrast, NSC processes are well documented and available at national level, with clear explanations for decisions made by the Department of Basic Education and the examinations quality assurance body, Umalusi, as indicated by its established rating (DBE, 2016).

Although the ratings allocated to systemic learning assessment programmes in the evaluation matrix in Table 8 may be debated, the following section summarises the overall findings in relation to the dimensions of effective systemic learning assessment arising from the literature. Lower than expected ratings frequently were attributed due to inadequate evidence or documentation for some of the sub-dimensions and dimensions of enabling context, system alignment or assessment quality. I argue that the presence of technical documentation is evidence of credibility and process integrity. Sample-based ANA was rated as emerging in terms of the technical aspects of assessment quality as the test results were not comparable from year to year (DBE, 2015a) though its curriculum alignment was very high as with NSC. The NSES was rated as between emerging and established, since it was carried out only once, as it was conceived of as a research project and therefore it included much more technical documentation on its implementation than the Systemic Evaluation which was rated as latent to emerging, despite using similar tests (Department of Education, 2003; 2005; 2008).

Using the original and modified quality standards for systemic learning programmes, international and regional assessments like TIMSS, PIRLS, and SACMEQ were rated as being technically sound and suited to measuring learning outcomes over time at the country level especially in the original framework. However, the curriculum alignment of these programmes needs monitoring and documentation. PIRLS is rated as lower in the original framework as the results and data, especially on African language pre-PIRLS performance in 2011 is not well known outside of the report

produced by the service provider which is available on the service provider's web site.

Modifications to Clarke's framework used in this study provide specific guidance on technical aspects of implementation which provide pointers to where technical expertise and best practice can be strengthened in the South African educational assessment system. The modifications are useful as they assist in identifying the strengths and weaknesses in certain sub-dimensions of the learning assessment programmes with a view to improving implementation at the country level.

In relation to assessment quality in the modified framework, test and item development and issues of technical validation, protocols and test content require attention across the board. Standardised international assessments such as PIRLS and TIMSS are consistently rated as established to advanced in these aspects of systemic learning assessment implementation at the country level in South Africa, particularly in respect of item and test development and design incorporating psychometric and bias checking among other mechanisms of improving technical assessment quality. Aspects of the security and protocols of the NSC are identified in the analysis as established and these may provide the basis for minimum standard for future assessment programme implementation and standardisation, while the established sampling methods used in the Systemic Evaluation from cycle to cycle and the National School Effectiveness Study (NSES) may provide insight into sampling decisions, and future analysis, reporting and dissemination efforts respectively.

The preceding analysis and chronology of assessments show evidence of South Africa's rich and diverse participation in systemic learning assessments. This encouraging state of affairs indicates the widespread awareness among education policymakers, researchers and practitioners of the benefits of judicious measurement of the quality of education at the country level. Some of the earlier learning assessments such as the MLA and NSES were once-off assessments which had weaknesses identified in the analysis. Even so, these once-off systemic learning

assessment programmes had technical dimensions which could help to strengthen the composition of future assessments.

Any systemic learning assessments developed in future should ideally reflect some of these characteristics and draw from the experiences of documentation and administrative integrity of these assessments especially those documented by independent providers. Sample-based ANA is certainly more standardised than Universal ANA although its weaknesses in terms of test design and coverage have been described in the literature (Spaull, 2013; Gravett and Henning, 2014). The value of sample-based ANA provides is that it is a curriculum-aligned assessment aimed at measuring learning outcomes in the lower grades, despite its limitations and emergent status. This provides information about instruction and learning progress as early as possible in learners' schooling careers.

The analysis shows that although, it is rated as emerging, sample-based ANA is well placed to form the basis of systemic assessment in South Africa provided it is modified using aspects of other systemic assessment that have been identified as technically strong and credible. ANA needs detailed technical modifications in qualitative, institutional and contextual terms. It requires better and more secure tests designed for comparability from year to year in addition to improved social and technical validation. Sample-based ANA requires better understanding of rational assessment system reform to be more broadly and generally understood. This means that the linkages between the assessment tools and instructional and learning improvement, uses and functions in the education system need to be more explicit in providing an enabling context and better system alignment in the assessment system.

It could be argued that the NSC, though stable, is inappropriate as a systemic learning assessment since it is a selective examination intended to certify completion of twelve years of schooling. It cannot be piloted as it needs to be confidential and it is not inclusive of all learners as many drop out of the system before Grade 12. Despite these limitations, the NSC results are a valuable indication of system

performance at the point of exit from the schooling system especially when adjusted for drop-out and combined with other indicators of learning performance.

The rating of the SACMEQ regional assessment which, though lacking in technical and analytical capacity, shows that it is established and well acknowledged as a source of information on learning progress in fifteen Africa education systems at regional level (UNESCO, 2015). The weaknesses identified in the SACMEQ programme require extensive investment in technical and specialist capabilities in respect of item and test development and adaptation over time, as well as bias checking, standardisation, documentation and technical analysis. Lack of technical documentation limits the replicability, reliability and ultimately compromises the integrity of any assessment programme and SACMEQ should therefore remediate this shortcoming urgently.

The modification of Clarke's framework and the SABER rubric carried out in this study is justified as it provides practical detail for where governments in developing countries such as South Africa should invest in order to more meaningfully measure the progress of learning outcomes. For example, supporting test and item development using psychometric means for to ensure test difficulty stability and equivalence would improve the low ratings attached to test design in all examinations and assessments in the system, provided that these efforts are rigorously carried out. Better test design, item development, calibration and piloting, and comparability are crucial in sustaining future systemic learning assessment efforts in South Africa. The importance of technical support in the areas of weakness identified in the analysis in this study cannot be overstated. The modifications recommended in this study may be incorporated into the assessment quality dimension of Clarke's framework as they provide further specification of areas which, if addressed, can contribute to strengthening the assessment system in South Africa, and more broadly, in the assessment systems of other developing countries in order to better track progress in learning outcomes.

Chapter 4 Review of past systemic learning assessment programmes in South Africa

This chapter presents a chronology and review of all efforts purporting to be systemic learning assessments used to measure progress in learning at the country level in post-Apartheid South Africa.

South Africa's extensive participation in educational assessment is reviewed in chronological order of implementation, ranging from long-standing programmes such as the National Senior Certificate (NSC) to the emerging Annual National Assessment (ANA) which started in 2011 are included along with the TIMSS and PIRLS international assessments, studies, surveys and data collections (such as MLA or SE) which provide insight into assessment implementation. The Western Cape Systemic Evaluation project was not included in this assessment as it is a sub-national programme, but warrants further research as it potentially holds lessons for measuring learning progress in the country. The chronology of assessments programmes in South Africa presented in Table 10, uses information extracted from technical reports, documentation and other literature available on assessment programmes in the public domain[64].

[64] ANA (DBE, 2010a, 2011, 2012b, 2013, 2015a), TIMSS (Reddy, 2006 & Reddy et al, 2012), MLA (DoE, 1999), SACMEQ (Moloi & Chetty, 2010; Ross et al, 2005), Systemic Evaluation (DoE, 2003b, 2003c, 2005), NSES (Taylor, N., 2011), PIRLS (Howie & van Staden, 2012; Howie et al, 2008) and NSC (DBE, 2014b, 2015c).

4.1 Chronology of systemic learning assessments in South Africa, 2016

The table below gives a profile and chronology of South Africa's extensive participation at the country level in systemic assessment since 1994. The table is presented as a prelude to the discussion of the attributes of the systemic assessment programmes in Chapter 4 in relation to enabling context, system alignment and assessment quality. The chronology of systemic assessments presented in the table includes the year, name, type of assessment, method of sampling, and grades and subjects tested along with background information collected using questionnaires administered to (P = pupils or learners, T = teachers, S = school principals, Par = Parents, Off = official dealing with the school, Cur - Curriculum). With respect to the subjects tested, EAL indicates that tests in English and Afrikaans were used. Superscript AfrL indicates that tests in indigenous African languages were used in addition to tests in English and Afrikaans.

Table 10: Chronology of systemic learning assessments in South Africa, 1994 to 2016

Year	*Programme*	*Type*	*Grade tested*	*Subjects tested*	*Pupils*	*Schools*	*Key SA analyses*
1995	Trends in Mathematics and Science Study (TIMSS)	International, sample-based, questionnaires for P, T, S and Cur	Grade 7/8 and Grade 12	Mathematics [EAL], Science[EAL].	9 792/14 000 including Grade 12 learners	251/400 including Grade 12 learners	Howie, S. & Pietersen, J.J. (2001). Mathematics literacy of final year students: South African realities. *Studies in Educational Evaluation*, 27: 7-25.
1999	TIMSS called TIMSS-R at the time	International, sample-based, questionnaires for P, T, S and Cur	Grade 8	Mathematics [EAL], Science [EAL], English language.	8 146	200	Howie, S. (2003). Conditions of schooling in South Africa and the effects on mathematics achievement. *Studies in Educational Evaluation*, 29: 227-241.
1999	Monitoring Learner Achievement (MLA)	International, sample-based, questionnaires for P, T, S and Cur	Grade 4.	Literacy [AfrL], Numeracy [AfrL], Life skills [AfrL].	10 759	400	Department of Education. (1999). *Report on the results of the Monitoring Learning Achievement (MLA) Project*. Commissioned by the Department of Education and supported by UNESCO and UNICEF. Author: J P Strauss, Research Institute for Education Planning, University of the Orange Free State. November 1999.

Year	*Programme*	*Type*	*Grade tested*	*Subjects tested*	*Pupils*	*Schools*	*Key SA analyses*
2000	Southern and Eastern African Consortium for Monitoring Education Quality (SACMEQ)	International, sample-based, questionnaires for P, T and S.	Grade 6	Language [EAL], Mathematics [EAL].	3 165	167	Moloi, M. Q. (2005, September). Mathematics achievement in South Africa: A comparison of the official curriculum with learner performance in the SACMEQ II Project. In SACMEQ International Invitational Conference, International Institute for Educational Planning (pp. 28-30)/ Van der Berg, S., & Louw, M. (2007). *Lessons learnt from SACMEQII: South African student performance in regional context*. University of Stellenbosch, Department of Economics and Bureau for Economic Research Working Paper, 16(07).
2001	Systemic Evaluation (SE)	National, sample-based, questionnaires for P, T, S, Par and Off	Grade 3	Literacy [AfrL], Numeracy [AfrL], Life skills [AfrL]	51 307	1 309	Department of Education. (2003b). Systemic Evaluation, Foundation Phase: Learners with Disabilities in Special Schools report also produced in 2003
2003	TIMSS	International, sample-based, questionnaires for P, T, S and Cur	Grade 8. Although Grade 9 learners sat the test as well.	Mathematics [EAL], Science [EAL].	8 952	255	Reddy, V. (2006). *Mathematics and science achievement at South African schools in TIMSS 2003*. Pretoria: HSRC. Available from: <http://www.hsrcpublishers.ac.za>
2004	Systemic Evaluation (SE)	National, sample-based, questionnaires for P, T, S, Par and Off.	Grade 6	Language [EAL], Mathematics [EAL], Natural science [EAL].	34 015	998	Department of Education (2005). *Grade 6 Systemic Evaluation: National.* Pretoria. Available from: <http://www.hsrc.ac.za/research/output/outputDocuments Gustafsson and Patel (2008).

Year	*Programme*	*Type*	*Grade tested*	*Subjects tested*	*Pupils*	*Schools*	*Key SA analyses*
2006	Progress in International Reading Literacy Study (PIRLS)	International, sample-based, questionnaires for P, T, S and Cur	Grade 4 normally but written by Grade 4 and 5 learners in SA.	Language [EAL and African languages– see footnote 65]	14 657 in Gr 5 and 16 057 in Gr 4	398 in Gr 5 and 432 Grade 4	Howie, S., Venter, E., Van Staden, S., Zimmerman, L., Long, C., Du Toit, C., et al. (2008). *PIRLS 2006 Summary Report: South African Children's Reading Literacy Achievement*. Pretoria: Centre for Evaluation and Assessment. University of Pretoria/Taylor, S., & Yu, D. (2009). *The importance of socio-economic status in determining educational achievement in South Africa*. Unpublished working paper (Economics) Stellenbosch: Stellenbosch University.
2007	Systemic Evaluation (SE)	National, sample-based, questionnaires for P, T, S, Par and Off	Grade 3	Literacy [AfrL] , Numeracy [AfrL], Life skills [AfrL]	54 449	2355	Department of Education (undated leaflet). 2008a. Systemic Evaluation: Grade 3 Literacy and Numeracy results.
2007	SACMEQ	International, sample-based, questionnaires for P, T and S	Grade 6	Language [EAL], Mathematics [EAL]. And HIVAIDS Knowledge Test [EAL]	9 071	392 (between 37 and 64 schools per province)	Moloi, M. Q., & Chetty, M. (2010). The SACMeQ III Project In South Africa/ Spaull, N. (2011a). A preliminary analysis of SACMEQ III South Africa. Stellenbosch: Stellenbosch University.
2007	National School Effectiveness Survey (NSES)	Longitudinal panel survey. Questionnaires for P, T and S.	Grade 3 (2007), followed through in Grade 4 (2008), followed through in Grade 5 (2009). First education panel	Language [EAL], Mathematics [EAL]	8 383 in each year of a 3 year panel	266	Taylor, N. (2011). *National School Effectiveness Study - Synthesis Report*. Johannesburg: JET Education. Taylor, N., Van der Berg, S. & Mabogoane, T. (2013). What makes schools effective? Report of the National Schools Effectiveness Study. Cape Town: Pearson.

[65] Achievement data for African languages was generally low with many missing values according to CEA Acting Head in August 2017.

Year	*Programme*	*Type*	*Grade tested*	*Subjects tested*	*Pupils*	*Schools*	*Key SA analyses*
2008	Annual National Assessments (ANA)	National, census (attempted), no questionnaires	Grades 1 to 6	Literacy [AfrL]/ language [AfrL], Numeracy [AfrL]/mathematics [AfrL]	Approx. 3.3m	Approx. 10,000	Trial run – no technical report published according to DBE (2011a).
2009	Annual National Assessments (ANA)	National, census with sample-based verification, no questionnaires	Grades 1 to 6	Literacy [AfrL]/language [AfrL], Numeracy [AfrL] /mathematics [AfrL]	Approx. 4.6m (sample approx. 12,700)	Approx. 14,000 (sample Approx. 510)	Trial run – no technical report published according to DBE (2011a).
2011	Annual National Assessments Universal (ANA - U)	National, census with no questionnaires administered.	Grades 1 to 6	Literacy [AfrL]/language [AfrL], Numeracy [AfrL] Mathematics [AfrL]	6 million	Approx. 20, 000	Department of Basic Education. (2011a). Report on the Annual National Assessments of 2011. Pretoria.
2011	Annual National Assessments Verification (ANA-V)	National, sample-based verification supplementing census collection. No background questionnaires.	Grade 3, 6	Literacy [AfrL] / Language [AfrL], Numeracy [AfrL]	129 375	1 800 (200 public schools per province)	Department of Basic Education. (2011a). Report on the Annual National Assessments of 2011. Pretoria.
2011	pre-PIRLS	International, sample-based, questionnaires for P, T, S, Cur and Par	Grade 4	Language [AfrL]	15 744	341	Howie, S., & van Staden, S. (2012). *South African Children's Reading Literacy Achievement - PIRLS and prePIRLS 2011 Summary of the key results (Media briefing).* Pretoria: Centre for Evaluation and Assessment.

Year	*Programme*	*Type*	*Grade tested*	*Subjects tested*	*Pupils*	*Schools*	*Key SA analyses*
2011	PIRLS	International, sample-based, questionnaires for P, T, S, Cur and Par	Sub-national sample of learners Grade 5 learners in English or Afrikaans Home Language (other countries Grade 4 learners participated)	Language[EAL]	3 515	92	Howie, S., & van Staden, S. (2012). *South African Children's Reading Literacy Achievement - PIRLS and prePIRLS 2011 Summary of the key results (Media briefing).* Pretoria: Centre for Evaluation and Assessment.
2011	TIMSS	International, sample-based, questionnaires for P, T, S and Cur	Grade 8 in other countries (but in 2011 written by only Grade 9 learners in SA)	Mathematics [EAL], Science [EAL].	11 969	285	Reddy, V., Prinsloo, C., Visser, M., Arends, F., Winnaar, L., Rogers, S.,….Ngema, M. (2012). *Highlights from TIMSS 2011: The South African perspective*. Pretoria, HSRC.
2012	Annual National Assessments Universal (ANA - U)	National, census with no questionnaires administered.	Grades 1 to 6 and Grade 9	Literacy [AfrL]/language [AfrL], Numeracy [AfrL] Mathematics [AfrL]	7.2 million	Approx. 24,000	Department of Basic Education. (2012b). Report on the Annual National Assessments 2012: Grades 1 to 6 & 9. Pretoria.

Year	*Programme*	*Type*	*Grade tested*	*Subjects tested*	*Pupils*	*Schools*	*Key SA analyses*
2013	Annual National Assessments (ANA-V) Verification	National, sample-based verification supplementing census collection. Background questionnaires for P, T and S.	Grade 3 , 6 and, in 2013, Grade 9	Literacy [AfrL] / language [AfrL], Numeracy [AfrL]	124 681	2 168	SAB&T Deloitte. 2013. Technical report on Verification ANA (V-ANA) results 2013 for the Department of Basic Education. Dated 11 February 2014. Unpublished.
2013	Annual National Assessments Universal (ANA - U)	National, census with no questionnaires administered.	Grades 1 to 6 and Grade 9	Literacy [AfrL]/language [AfrL], Numeracy [AfrL] Mathematics [AfrL]	7 million	Approx. 24,000	Department of Basic Education. (2013). Report on the Annual National Assessments 2013: Grades 1 to 6 & 9. Pretoria.
2014	Annual National Assessments Universal (ANA - U)	National, census with no questionnaires administered.	Grades 1 to 6 and Grade 9	Literacy [AfrL]/language [AfrL], Numeracy [AfrL] Mathematics [AfrL]	7.4 million	Approx. 24,400	Department of Basic Education. (2014a). Report on the Annual National Assessments 2014: Grades 1 to 6 & 9. Pretoria.
2015	TIMSS	International, sample-based, questionnaires for P, T, S and Cur	Grade 9	Mathematics [EAL], Science [EAL] .	12 514	292 schools, 334 Mathematics and 331 Science teachers.	Reddy, V., Visser, M., Winnaar, L., Arends, F., Juan, A and Prinsloo, C.H. (2016). TIMSS 2015: Highlights of Mathematics and Science Achievement of Grade 9 South African Learners. Human Sciences Research Council.

Year	*Programme*	*Type*	*Grade tested*	*Subjects tested*	*Pupils*	*Schools*	*Key SA analyses*
2015	TIMSS-Numeracy (TIMSS-N)	International, sample-based, questionnaires for P, T, S, Cur and Par.	Grade 5	Mathematics [EAL], Science [EAL] .	10 932	297 schools, 10 500 learners' parents or care givers and 297 maths educators.	Reddy, V., Visser, M., Winnaar, L., Arends, F., Juan, A and Prinsloo, C.H. (2016). TIMSS 2015: Highlights of Mathematics and Science Achievement of Grade 5 South African Learners. Human Sciences Research Council.
Annual	National Senior Certificate (also known as Matric)	National, universal/ census, all students in Grade 12. No questionnaires.	Grade 12	Approximately 130 subjects in 2016 [EAL] Language[AfrL]	Approximately 674 000	Approx 6 767 centres	Department of Basic Education. (2016). 2016 National Senior Certificate Examination Report. Pretoria: Department of Basic Education.

4.2 Description of systemic learning assessments in South Africa, 2016

The narrative review presented here is structured using the three dimensions of the modified version of Clarke's framework: namely the enabling context; system alignment; and, assessment quality of the assessment programmes analysed and evaluated in the previous chapter. The systemic assessment programmes are presented in chronological order with a summary of findings arising from the literature in the chronology of systemic learning assessments in South Africa in section 3.1 above.

4.2.1 Enabling context

This section examines the enabling context for the implementation of systemic learning assessment programmes since 1994 in South Africa. It examines the following sub-dimensions of such programmes: social validation; institutional and policy context; sustainable funding; and integrity of organisational structure, capacity and human resources.

The National Senior Certificate (NSC) is an established examination which certifies the completion of twelve grades of schooling, and allows admission to higher education institution programmes and the world of work in South Africa (DBE, 2014b, 2015c). The yearly NSC examination results are widely debated in academic, research and public circles as an indicator of the health of the schooling system. Before 1994, several systems and examination-setting processes existed for the exit qualification at Grade 12 with variable consistency and quality assurance arrangements within the many education systems. Umalusi, an independent quality assurance body, sets standards in the NSC certification process and was instrumental in standardising the NSC in 2008 (DBE, 2016). The NSC has been standardised with nationally set common papers and a high security administration process which is documented, audited, well resourced, funded and coordinated through nine

provincial education departments by the National Department of Basic Education, supported by Umalusi – an arrangement which some feel should include a separate examination agency in the long run (Poliah, 2014; DBE, 2016). The importance attached to the examination results is sometimes inappropriately high, prompting moves to develop baskets of indicators of school performance incorporating examination and assessment performance adjusted for efficiency at Grades 3, 6, 9 and 12 (DBE, 2003; DBE, 2016).

The Trends in Mathematics and Science Study (TIMSS) is an established international assessment which was first implemented in South Africa in 1995 among Grade 8 learners, although in 2002, both Grade 8 and Grade 9 learners participated (Reddy, 2006 & Reddy et al, 2012). Consistently low performance from Grade 8 learners in TIMSS in developing countries in 2002 led to the decision to test only Grade 9 learners using the Grade 8 test in Botswana, Honduras and South Africa, among the few developing countries who participated in the study. The Human Sciences Research Council (HSRC) carries out TIMSS on behalf of the country, funded through the national Department of Basic Education with coordination and oversight by the DBE Examinations unit. The TIMSS reports are released by the Minister of Basic Education and are used to benchmark and to triangulate information from local assessments and from other sources of data about education quality[66]. Training and capacity building is concentrated mainly within the structures of the service provider, with some research workshops convened after the release of each cycle of data (Reddy et al, 2012).

The United Nations Education, Scientific and Cultural Organisation (UNESCO) and United Nations Children's Fund (UNICEF) sponsored the Monitoring Learning Achievement (MLA) initiative which was set up to monitor the EFA goals in developing countries through a survey of basic learning competencies in 1999 (Chinapah, H'ddigui, Kanjee, Falayajo, Fomba, Hamissou, et al, 2000). The MLA was a once-off research and reporting exercise and is therefore rated as latent, despite widespread political buy in through the global EFA process. The MLA report

[66] Basic Education Director-General's remarks at the Basic Education Sector Lekgotla held on 23 to 25 January 2017.

preparation in South Africa was supported with funding from UNESCO and UNICEF. Additional workshops and support were also provided for research practitioners and education officials (DBE, 1999) as the MLA project outcomes included training and skills development in the assessment, monitoring and analysis of performance data. Provincial officials assisted in the administration of the MLA coordinated by the Quality Assurance Unit in the national Department of Education while the survey, collection and analysis was carried out by the Research Institute for Education Planning on behalf of the Department of Basic Education. The results were used for planning, resource allocation and understanding of resource disparities in the post-Apartheid education system (DBE, 1999).

Southern and Eastern African Consortium for Monitoring Education Quality (SACMEQ) is a well-known, fifteen-country regional assessment in Southern and Eastern Africa rated as established (Moloi & Chetty, 2010; Ross et al, 2005). Its origins date back to 1991 when several Ministries in the region began to work with the International Institute for Education Planning at UNESCO. Currently, the SACMEQ coordinating unit is based in Botswana, having been based in Paris until early in 2015. A Ministerial Steering Committee which meets every two years, guide the project and fund the implementation which is driven by national research co-ordinators stationed within ministries, although competing priorities sometimes mean that data collection and analysis are delayed. The capacity development aspect of SACMEQ is prominent and data collection is institutionalised within ministries using officials (Gustafsson and Moloi, 2011). Policy issues of interest such as HIV/AIDS and TB knowledge are included to ensure that SACMEQ is relevant to the needs of all fifteen participating countries. SACMEQ is well respected as a source of comparative information on education systems in the developing world (Best et al, 2013; UNESCO, 2015).

The Systemic Evaluation (SE) is rated as latent mainly due to its limited number of cycles (DoE, 2003b, 2003c, 2005). SE was identified in the Assessment Policy for the General Education and Training Band (1998) and was carried out in Grade 3 in 2001, Grade 6 in 2004 and Grade 3 again in 2007 using a sample-based evaluation

instrument piloted in 2000. SE provided a baseline of learning outcomes in a collaborative effort co-ordinated by the Department of Basic Education and was carried out by a consortium of service providers at the Research Institute for Education Planning, the Centre for Education Policy and Development, the HSRC and provincial education departments, with the latter involved in co-ordination, collection and scoring. This supplemented the National Policy on Whole School Evaluation (2001), which governed school processes in support of school improvement (DoE, 2001).

The Progress in International Reading Literacy Study (PIRLS) focuses on the purposes of reading (literary experience and acquisition of information accounts for half of the items), processes for comprehension and reading attitudes and behaviour of learners (Howie and van Staden, 2012). In South Africa, PIRLS assesses reading literacy and comprehension for children with different home languages. PIRLS defines reading literacy as the ability to understand and use written language valued by individuals and required by society. In order to guide investments in literacy and reading, PIRLS explores the amount of reading, the value attached to it and how children are taught to read. Progress in International Reading Literacy Study (PIRLS) is an established standardised assessment with protocols and guidelines for administration, analysis, dissemination and reporting, developed and monitored by the IEA (Greaney and Kellaghan, 2008; Howie et al, 2008; Howie & van Staden, 2012). Pre-PIRLS was developed for countries exhibiting low performance in the original PIRLS assessment by the IEA and provides and assessment of basic reading skills that are prerequisites for success on PIRLS and was implemented in 2011 in South Africa among Grade 4 learners following the low performance in PIRLS 2006 among learners of that grade. PIRLS 2006 and PIRLS 2011 and prePIRLS 2011 tests for reading and literary were also administered in Grade 4 and 5 in South Africa. Contextual questionnaires and all instruments except test instruments were developed in English. All tests were developed by the IEA and translated into South African Languages as needed. Reading literacy passages were translated to the 11 official national languages.

PIRLS and pre-PIRLS was the only standardised international assessment of reading and literacy in all official languages in South Africa and was last reported in 2011, although the 2016 PIRLS cycle was underway at the time of writing. All PIRLS assessments were run on behalf of the Department of Education under the auspices of the Centre for Education Assessment at the University of Pretoria, with restricted access to African language item-level data (see footnote 47). This lack of access has stifled the use of the data set, further analysis, and the development of research and knowledge from a standardised assessment of literacy in the African languages which is technically credible and can be benchmarked. PIRLS findings recently became more well known in education policy debates due to various high level pronouncements on the barriers to literacy and reading among young children and adults, by the Minister of Basic Education in her Budget Speech in 2015[67].

Box 2: PIRLS Comparisons possible across years and groups (information provided by Acting Director, Centre for Educational Assessment, Pretoria. August 2017)

	2006, 2011 & 2016	2011 & 2016	2006 & 2016
Grade 5	Achievement & questionnaire data for English and Afrikaans (benchmark data)		Achievement & questionnaire data for English, Afrikaans & isiZulu. Language groups **or** provinces can be compared*
Grade 4	Questionnaire data for all 11 languages (not achievement due to low performance and missing data) Questionnaire data for all provinces*	Achievement and questionnaire data for all 11 languages (not provinces).	

*Languages cannot be compared within provinces. Language groups and provinces should be compared separately

[67] https://www.gov.za/speeches/minister-angie-motshekga-basic-education-dept-budget-vote-201516-6-may-2015-0000

The National School Effectiveness Study (NSES) is rated as latent due to its limited number of cycles (Taylor, N., 2011). It was carried out between 2007 and 2009 and was the first large-scale panel study of educational achievement in South African primary schools, tracking school management and processes over three years in a nationally representative panel of primary schools. The NSES, related home-, school- and classroom-level factors to student learning in South Africa, using literacy and numeracy tests administered in a nationally representative panel study according to Taylor, N. (2011) and Taylor, Van der Berg, & Mabogoane (2013).

The NSES design allowed for calculation of the amount of learning occurring within a year of schooling in South Africa. The multi-partner implementing consortium for the project was managed by JET Education Services and funded by the Royal Netherlands Embassy, and involved academics from universities and research agencies working on independent projects related to the determinants of quality of learning outcomes. The national Department of Basic Education coordinated the study in eight out of nine provinces. Gauteng did not participate as it was administering another assessment at the time of the study (Taylor, S., 2011). The policy context did not require that the survey be institutionalised although many aspects of the monitoring were carried through into later studies on schools including the 2011 School Monitoring Survey (DBE, 2015a) which monitored school- and principal-support activities initiated at district level.

The Annual National Assessment (ANA) was launched after a Presidential announcement of annual testing in Grades 3, 6 and 9 in the 2011 State of the National Address (Zuma, 2011), following two trial implementation cycles in 2008 and 2009 (DBE, 2011a). The enabling context of sample-based ANA is rated as emerging to established, and it was in its early stages of development when it was halted in 2015 due to opposition from teacher unions. ANA was initially strongly supported by unions and other stakeholders, and was accompanied by a massive workbook programme, the launch of the newly specified national Curriculum Assessment Policy Statements (CAPS) and two versions of assessment in a reform package: a universal ANA for instructional improvement, and a sample-based ANA

for monitoring learning outcomes at system level (DBE, 2010a and 2015b). Despite parental approval in the media, antipathy against universal testing began to be expressed in the system, especially against the additional administration and expectations imposed on teachers by the ANA implementation requirements at school, district and provincial level (SADTU, 2014). Social validation and capacity for implementation of the different assessment tools associated with the ANA was weak and this contributed to the abrupt end of the ANAs in 2015.

In summary, the extensive participation in learning assessments over the last two decades indicates a healthy recognition of the importance of information about learning outcomes in South Africa's education system. The NSC, TIMSS, PIRLS, SACMEQ and sample-based ANA all enjoy a positive enabling context while the MLA, NSES and SE have low ratings as they were *ad hoc* programmes which are no longer implemented. Although there is evidence of tangible political and administrative support and resourcing for educational assessments, it could be argued that the development of the assessment system has not kept pace with that in curriculum reform over the last two decades. This lag imposes limitations on methods for ensuring the social and technical validation of learning assessments. With respect to technical validation, there is little evidence of research into the use, understanding or perceptions of different forms of assessment, despite the considerable resources allocated to the assessments and the education system more broadly. In addition, the service provider relationships of implementation need to be broadened to include more technical advice sourced for specific parts of the assessment implementation to enhance and validate technical decisions made in the administration of the assessments. Creating an enabling context for an effective systemic learning assessment system therefore requires a better understanding of the dynamics, perceptions and nature of the social and technical validation networks at different levels in the system. This requires careful engagement, information and research to inform communication and system development, in addition to the alignment of assessments in the education system.

4.2.2 System alignment

Following the review of the enabling contest of different systemic assessments, this section examines the system alignment for the implementation of systemic learning assessment programmes since 1994 in South Africa. The sub-dimensions relating to the extent of curriculum alignment of learning assessments and teachers' understanding of learning assessment are examined. The modified sub-dimensions include expanding understanding to other partners and officials in the education system. The components of an effective systemic assessment system must be aligned in form, function and focus with other assessments in the broader education system, in support of curriculum mastery (Darling-Hammond and Wentworth, 2010).

The NSC examinations are established in terms of both sub-dimensions of system alignment and the subjects examined are closely curriculum-aligned to CAPS as it is an examination for certification at the end of Grade 12 (DBE, 2016). Over 40 000 markers, drawn from the teaching corps with relevant subject based experience through all nine provinces, are involved in marking over 10 million scripts. Familiarity and teacher remuneration are therefore taken as given. Although the NSC in its current form has only been in existence for eight years at the time of writing, there have been other examinations at the end of schooling for many decades in South Africa. Teachers and all stakeholders understand the assessment and it holds educational and public attention through the media.

TIMSS is rated as emerging to established for this dimension as it has some weaknesses in the extent of its curriculum alignment and understanding of the assessment by teachers and others in the education system. Online documentation on TIMSS (www.timss.bc.edu) shows that only around 20 per cent of the assessment items (22% in 2002) overlapped with national curricula according to a Test Curriculum Matching Assessment (TCMA) for earlier versions of TIMSS. In TIMSS 2011, about a fifth of the test items matched the taught curriculum in South Africa (Reddy et al, 2012) although this proportion reportedly increased in the 2015

TIMSS[68] due to the improved match between the national curriculum topics and those tested in the TIMSS assessment framework. The TIMSS test was set in Afrikaans or English as these are the two languages of learning and teaching in the Grade. TIMSS results are slowly being incorporated into sectoral reporting narratives although it is an assessment which is not well known by researchers, practitioners or policy makers at provincial, district and local level through advocacy and communication exercises such as the launch of a TIMSS-SA website and roadshows with the Department of Basic Education in 2016. Teachers' understanding of TIMSS needs to be expanded. The advantages of TIMSS are that it tests educators as well as learners, and background information and data are collected for learners and teachers.

MLA was only carried out once so system alignment is rated as latent (DOE, 1999). Instruments were developed and piloted in a number of Southern African countries following the instrument development. Principal, learner and educator questionnaires were also administered for background information. Curriculum alignment is not clear from the report. The instruments were reportedly developed in consultation between the South African implementation team, translated into the 11 official languages in consultation with UNESCO in relation to the basic education competencies in literacy, life skills and numeracy accepted as a minimum (DOE, 1999). Instruments and items could not be traced at the time of writing, although according to the South African report, country teams consulted with curriculum experts. No evidence of teachers' understanding of the assessment was available. The external reporting intent was explicit in all MLA reporting in response to global education commitments.

The SACMEQ is established and curriculum alignment happens in every cycle according to reports from the National Research Coordinator (NRC) in South Africa[69]. According to the national coordinator, the primary purpose of the SACMEQ was to expand opportunities for educational planners to gain technical skills required to

[68] This was reported by Dr V Reddy of the HSRC during the presentation of the 2015 TIMSS results for Grade 5 and Grade 9 in November 2016.
[69] Dr Mark Chetty, DBE, 2016

monitor and evaluate the general conditions of schooling and the quality of basic education in their respective systems. Over the years since the first project in 1995, SACMEQ has developed research instruments and collected useful information using advanced research methods. The curriculum of each country is assessed against the items and was declared to be aligned over a decade ago (Ross et al, 2005) and this has been reconfirmed in country reports (Moloi and Chetty, 2010). SACMEQ was designed to generate valid measures of levels and changes in achievement (a) across countries at single time points and (b) across time points for individual countries. To achieve this goal, SACMEQ reportedly follows almost the same methodologies across studies and uses the same instruments which must be kept confidential to remain valid. The methodology and instruments used in the SACMEQ IV project in 2013 were reportedly similar to those in SACMEQ III according to the SACMEQ Coordinating Centre in 2016. The curriculum relevance of the SACMEQ survey is illustrated by the inclusion of an HIV and AIDS knowledge test (HAKT) of Grade 6 learners and their teachers. In the SACMEQ IV project, TB knowledge levels of learners and teachers were also tested in addition to the HAKT. Workshops have been convened using the SACMEQ data in the first three cycles by various academic and research agencies as the comparative data provides opportunities for understanding education progress in the participating countries (Moloi and Chetty, 2010).

The latent rating of the Systemic Evaluation is due to the limited number of cycles of the evaluation completed in 2001, 2004 and 2007. High level strategic reporting was supported by leadership in education, and was carried out in the Systemic Evaluation at provincial level. Shortages were the predominant focus of much of the reporting (DBE, 2003b, 2003c, 2005). Instruments were drafted by education specialists, stakeholders and Department of Education officials, with benchmarking of the tests by four specialists assigned to the task. Translation was done in-house and instruments selected based on difficulty, bias and discrimination indices which did not use item response theory.

The lower than expected rating of emerging for PIRLS is mainly due to lack of evidence of awareness-raising or advocacy events at the country level, aimed at a broader audience of teachers, researchers or planners in the sector. PrePIRLS and PIRLS item level data in the African Languages were not available, even on request, to support language literacy efforts in the foundation phase in 2016 (see footnote 47). As such, prePIRLS has not been fully exploited for decision and policy support in the education system. Considering the reading and literacy constraints and concerns in the foundation phase in African languages in the country (Spaull, 2013; DBE, 2015a), this is regrettable.

A low rating of latent was allocated to the NSES as it was implemented once as a research study. According to Taylor et al, (2013), students in 266 schools in eight of South Africa's nine provinces were tested in language and numeracy in the NSES in 2007 (Grade 3), 2008 (Grade 4) and 2009 (Grade 5). The 2007 sample was a sub-set of the Systemic Evaluation sample used in Grade 3 and provided the opportunity to assess learner performance in English with mother tongue tests administered in the same year. The same learners were tested in each year and a valuable panel dataset was constructed including academic achievement and other information about approximately 8 000 learners across three years. In addition to information on home language competence, background information was collected using asset-based measures rather than reported parental education and family income, as younger children were covered in the study. Grade 3 Systemic Evaluation tests aligned to the curriculum were used. They were administered only in English as most schools for African children change their medium of instruction in Grade 4 from mother tongue to English. The same tests were administered each year, making the results comparable from one year to the next. Background information on the ex-racial allocation of the school was populated using the Department of Education's Master List of Schools. While at Grade 3 level the children would have been disadvantaged by writing in a language with which they were unfamiliar, this design enabled the comparison of scores directly across the three years. Also, because the NSES schools were a subsample of the Systemic Evaluation sample, the design provided a unique opportunity to compare scores by the same children on the same test written in a

different language. The use of an unchanged Grade 3 test may have placed a ceiling on the gain scores of more able learners at the top end of the score distribution although it was found that few learners fell into this category as any findings were unlikely to have been skewed by a few high achievers (Taylor, N., 2011).

According to Taylor, N. (2011), the NSES cohort design enabled gain scores of learners to be related to the characteristics of the learner's teacher in the year, while controlling for school and home variables. A teacher test was administered but was found to be too limited in scope to make conclusions possible, and the SACMEQ 2007 survey teacher test results were used instead to draw conclusions about teacher knowledge (Taylor, N., 2011).

Sample-based ANA is rated as tending towards established as test development was done by departmental officials working with external experts. Assessment frameworks were developed using curriculum standards and expectations at the time, and ANA is highly curriculum aligned with the curriculum standards at the country level. Since this dimension deals with links between the assessment and learning standards, it is limited, and needs to be supplemented by technical quality dimension ratings. Since there was little evidence for comparability in test design or item difficulty between years (Cartwright, 2013). Piloting was done in 8 schools in 2011, and teacher and learners were interviewed about the perceived difficulty of the tests. Teacher participation was secured, with unions and principals were mainly required to oversee administration at school level as was the case with the Grade 12 examination. This limited the potential for standardisation as schools were required to support learners in Grades 1 to 3 by reading out questions. Teachers marked the tests according to protocols in universal assessment but their assessment knowledge was not explicitly quantified in the process. Documentation on decisions in relation to scoring and analysis in the ANA process is detailed in some service provider technical reports (SAB&T, 2013) but documentation on decisions and changes in test administration, test development and processing is not detailed enough over time s to enable tracking of changes in implementation and administration over time (DBE, 2011a; DBE, 2014a).

System alignment is more difficult to achieve in international assessments and is more likely to be achieved through national learning assessment programmes as these are more likely to be aligned to curriculum and learning standards at the country level. In this analysis, the NSC, SACMEQ and sample-based ANA are most well-known and aligned to the national curriculum. However, curriculum alignment is one aspect of system alignment. Sample-based ANA is highly rated is related to the exposure of the ANA tests to thousands of teachers in the system. This confirms the appetite for the measurement of learning outcomes in the early grades despite the unintended negative consequences related to social validation and technical dimensions of the implementation of the assessment. Learners, teachers, officials, parents and the general public should understand the assessments and their role, nature and functions in a fully effective assessment system. A recommendation in this study is that to enhance system alignment, these role players and partners require a technical understanding as well as increasingly broad based educational understanding of the interrelationship of different assessments. Already some of this work is being done – and training and capacity building efforts are becoming institutionalised in the TIMSS and NSC cycles, and even the work done in developing templates for ANA reporting and diagnostic reporting from the universal ANA results are noteworthy. A richer and broader understanding of assessment tools, assessment functions and linkages to other parts of the education system and processes can help to support curriculum mastery, education reform and tracking of learning improvement and progress.

4.2.3 Assessment quality in the original framework by Clarke (2012b)

This section examines the assessment quality of the implementation of systemic learning assessment programmes in post-apartheid South Africa. It examines the following sub-dimensions of programmes according to Clarke's original framework: inclusiveness of the assessment; use of assessment data; considerations in reporting and dissemination; the presence of technical documentation; and the use of assessment results in knowledge management. The modifications to the framework

in the next section resulted in slightly different evaluations of the different assessment programmes based on their technical characteristics.

The NSC is established and is taken by all learners who reach Grade 12 for certification so the need for sampling documentation is absent. Accommodation for learners with special needs are made in the assessment and these are strictly monitored for blind learners (braille-versioned), deaf learners and learners who need reading and writing accommodations in the examination process. There is an endorsed NSC for learners with special education needs with dyscalculia, aphasia and dyslexia (DBE, 2016). Diagnostic reports based on a sample of problems faced by learners arising from the marking, scoring and moderating processes are produced for the 11 subjects with highest enrolment to inform curriculum provisioning and process fidelity. Irregularities are identified and reported while historical data on learner performance for five years is used to determine a norm against which current performance is compared in a process approved by Umalusi, the quality assurance body for the NSC certificate. Assessment administration and security concerns are the focus of NSC process standardisation. This requires a focused continuous NSC quality assurance, research and evaluation programme for all aspects of administration and a separate focus on content, administration, analysis and reporting.

TIMSS implementation is established in country. TIMSS 2011 used a matrix-sampling technique that spread the test burden across learners using 12 test booklets. Oversampling allowed conclusions to be drawn down to provincial level. The School Register of Needs database was used as the sampling frame by province and language of teaching in a three-stage sample (Reddy et al, 2012). Information on teachers and their preparation, classroom characteristics and school context information was provided by teachers and principals. Learner background information was provided by learners. The unit of analysis is the learner in cases where information from the school and teacher questionnaire was used. Curriculum context was provided in a questionnaire and the responses from countries were used to provide an overview of the curriculum landscape in the country for a TIMSS

encyclopaedia of curriculum arrangements (Mullis, Martin, Minnich, Tanco, Arora, Centurino & Castle, (2012). Sampling was done by taking a probability proportionate to size approach (PPS) with all learners in one intact class in one grade chosen up to a maximum of 40 learners. Advanced statistics and Item Response Theory were used in the analysis. Documentation on scoring, analysis, collection, survey methods, data cleaning, data capture, item piloting and item development was used to standardise processes in 2011 (Mullis et al, 2012) and 2015 (Reddy et al, 2015).

The latent rating of MLA is due to its lack of documentation and specifications on methodology and preparations apart from schools sampled (DoE, 1999). Schools with fewer than 30 learners in the MLA study for Grade 4 were excluded and 400 schools were selected proportionally from the provinces depending on the numbers of schools per province, with 30 Grade 4 learners chosen per school. Data collectors and field workers were trained according to a manual. The instruments were in all eleven official languages, with English background questionnaires on schooling context, learner background and teaching context. Analysis and reporting were produced on literacy and numeracy tasks in quartiles using percentages of correct items and average performance. Results were reported by province, gender, school location and domains of tasks assessed. Profiles of school and teaching resources were produced per province including teacher provisioning, equipment and furniture, participation and community support, and teacher and instructional aspects of the schools. Other aspects of schooling examined included school management, governance, provisioning of materials, educator appraisal and development, school endowment, parental participation, learner assessment, educator absenteeism, educator absenteeism and perceptions and career aspirations.

SACMEQ is rated as tending to established although it has technical weaknesses in relation to documentation and standardisation of administration reflected in the rating allocated to it in the modified and more technically aligned framework. SACMEQ country samples were drawn in order to yield standard errors of sampling for learners in Grade 6 such that a sample estimate such as a ‘mean’ of a population

percentage would have a standard error (SE) of ±2.5 percent.[70] However, for the SACMEQ III study, documentation was not available on the extent of curriculum alignment or on different aspects of the methodology or treatment of raw scores specifically for the 2007 SACMEQ assessment. Through the use of the Rasch model for psychometric analysis of item-level test performance, SACMEQ reports performance in hierarchies of skills and knowledge that learners and teachers demonstrate in the tests. Eight levels of achievement are used (Moloi and Chetty, 2010).

The sampling frame for the SACMEQ study was obtained from the EMIS database in 2013 and submitted to the SACMEQ Coordinating Centre (SCC). The SCC used expert statisticians to train national research coordinators and their deputies to draw samples for each of the member countries according to a presentation on the SACMEQ results at the Department of Basic Education, Pretoria on 18 August 2017.[71] The SACMEQ IV survey in South Africa in August 2013 yielded information collected from a total of 7 046 learners, 775 teachers and 298 school heads across all nine provinces of the country. The survey involved testing of a sample of Grade 6 learners and their teachers in Reading and Mathematics, HIV/AIDS knowledge and collecting contextual data through administering specially-designed self-completed questionnaires. The focus of the questionnaires was on the conditions influencing teaching and learning in schools.

[70] It is important that each statistic such as the 'mean' is interpreted in association with its sampling error. For this level of sampling accuracy, it is possible to be sure 19 times out of 20 that the population value of a given percentage lies within ±5 percent (±2 x 2,5% = ±5%) of the estimate derived from the sample. For example, if the sample estimate of female learners in Grade 6 is 49%, then it can be claimed with 95% confidence that the mean percentage of female Grade 6 learners in the population will be 49%±5% which ranges between 47.8% and 50.3%.

[71] Where the number of Grade 6 learners was 25 or less than 25 in a school, all the Grade 6 learners were included in the sample except where the number of registered Grade 6 learners was less than 15. The "excluded" population of learners was 4.7 percent, which was slightly less than the stipulated 5 percent to meet the SACMEQ criteria for accuracy in large-scale assessment data. A two-stage sampling design was used. In the first stage, schools in the "defined" target population were sampled on a "probability-proportional-to-size" (PPS) basis. The PPS sampling technique meant that relatively large schools had a higher probability of being selected than smaller schools. In the second stage, learners were sampled from all the Grade 6 classes in each of the sampled schools using the SAMDEM computer programme (IIEPSAMP_V1.3). Twenty five learners (minimum cluster size) were sampled where the total number of all enrolled Grade 6 learners at the time of data collection was greater than 25.

Systemic Evaluation is rated as latent in terms of assessment quality, mainly because of the limited documentation, standardisation and a lack of documentation on empirical methods used to analyse performance and learning changes in the programme. Sampling used the Schools Register of Needs list of schools as the sampling frame for a 5% sample in the Systemic Evaluation. The sample was stratified by urban and rural milieu and included all districts and farms schools in order to report on district level resource endowments. Items and instruments for the 2007 Systemic Evaluation effort and the sample-based ANA 2011 were identical, although they had design limitations arising from inappropriate test design skills levels (Howie, Barry and de Kock, 2011). All schools were arranged by province and district and a random selection of schools made, with at least 30 per class in the grade selected for sampling. Schools were chosen randomly until the 5% of learners in the sampled grade was reached. Data collection was carried out by trained departmental officials trained and scoring was carried out by district staff, with moderation of 5% of scripts. Data capture, scoring and verification was done by a consortium of providers. Analysis in the Systemic Evaluation focused on percentage pass differences by gender and on the distribution and levels of resources in the system. Reports emphasised school management, teacher training, practice and attendance, in addition to stakeholder perceptions and materials and resource provisioning. Factors influencing performance were identified, including safety and security, quality of schooling context, resources for teaching and learner background. Achievement levels were arbitrarily assigned in terms of percentage pass levels rather than performance levels (DoE, 2003b, 2003c, 2005).

The Progress in International Reading Literacy Study (PIRLS) is rated as emerging in relation to its country level implementation despite the advanced design and standardised implementation protocols overseen by the IEA internationally (Howie & van Staden, 2012; Howie et al, 2008). This is because Clarke's original framework emphasises the application, use and dissemination of the results in support of policy and practice in its sub-dimensions. In the years before 2016, workshops or opportunities to learn about the data were not widely available outside of the service provider and selected researchers, and rare opportunities were presented at

country level understand the data outside of the implementing agency (see footnote 47). PIRLS is largely not widely known in the education system[72] and this affects its rating in the evaluation in this study. The main problem with the PIRLS in South Africa is difficulty of access to item level responses in African languages for any two cycles. In 2016, the head of the unit tasked to implement the PIRLS study was not in the position to provide this information despite being requested officially for learner performance information at the learner and item level in all African languages tested in the PIRLS 2006 and pre-PIRLS 2011 and PIRLS 2011 cycles (see footnote 47) as appropriate.

The NSES programme is rated as emerging to established since it was a research study with well documented, standardised procedures, with results which were produced using advanced statistical analyses (Taylor, N., 2011). The test design incorporated a range of grade-appropriate questions (originally developed for systemic evaluation) and the item difficulty level of test questions used ranged from Grade 1 level to Grade 4 level. Reporting was thus mainly done using average and gain scores achieved by learners. The mean achievement in literacy in 2007 (Grade 3) was 19%. This improved to 27% in the following year. For numeracy, the mean achievement increased from 28% in Grade 3 to 35% in Grade 4. Reporting was in terms of gain scores, frequency distributions of score averages by socio-economic status, historical school type and functionality, along with provincial, gender and population group disaggregation. Average performance was reported in addition to item-level test performance. Learning gains were found to be inappropriately low, with learners demonstrating too little conceptual knowledge to achieve learning outcomes in the NSES. Technical documentation was freely available from the service provider and the NSES informed the development the education sector plan in 2010 and contributions to the National Development Plan (NPC, 2012).

The ANA in South Africa is an emerging learning assessment programme in a country with a national curriculum framework. There were two versions of ANA: a universal

[72] A poll of district managers in a training programme on data analysis in 2016 the majority were unaware of the PIRLS programme in South Africa apart from what was contained in the Ministerial pronouncements.

.

version and a separate sample-based independently verified systemic assessment version called Verification ANA (DBE, 2011a), though the whole ANA programme was brought to a halt just before the 2015 was to have been written. Verification ANA and universal ANA differ in that at the last the latter is a universally administered to all learners in Grade 1 to 6 and 9 for the purposes of diagnosis and the former a sample-based systemic assessment administered by independent service providers. Verification ANA is referred to as sample-based ANA and was analysed in this study despite its shortcomings which include: (i) the lack of comparability across years, as it has no anchor items in the test and (ii) tests have not been kept secure as the same test used in sample-based ANA has been administered universally in an effort to improve diagnosis at school level. This is a design fault which needs to be rectified in future reforms to the assessment system, and which influences its rating as an emerging assessment programme. Sample-based ANA is rated as an emerging systemic assessment. District officials trained by the independent service provider ultimately bore responsibility for ensuring standardised administration (DBE, 2011a. 2012b, 2013, 2015a) which may not have been perfectly implemented given the later emerging accountability pressures at school level alluded to in the critique of the ANA programme. Various steps were taken to improve the diagnostic utility of the ANA including scoring and marking of a sample of the scripts in each grade and the analysis to inform curriculum and assessment personnel[73] on potential areas for improvement.

The original assessment quality dimension includes aspects of the use, utility and considerations in the reporting, application and dissemination of the assessments as well as knowledge management and the modified framework including technical elements of assessment. The modified framework emphasizes the technical security and confidentiality concerns, the standardisation of learning assessments and the reduction of bias in results, as well as technical issues of item, test and sampling design, as well as reporting of test score growth and performance levels. Both allow the analysis and the identification of strengths and weaknesses of programmes for

[73] Tests were versioned for Grades 1 to 3 into 10 languages in addition to English and adapted for blind, short-sighted and deaf learners for universal version of ANA.

measuring learning progress at the country level, although the modified framework focuses more on the more technical sub-dimensions of assessment programmes for measuring learning outcome trends in South Africa[74].

4.2.4 Assessment quality (modifications and added technical sub-dimensions)

This section reviews the systemic learning assessment programmes in South Africa which are analysed against the modified assessment quality sub-dimensions of Clarke's framework and the SABER rubric (2014) which arise from the study findings and literature. The modifications made are mainly of a technical nature in the assessment quality dimension of the framework and rubric. The modifications include the following sub-dimensions: Security and confidentiality concerns; Standardisation of learning assessments to reduce bias; Test design and development; Analysis of assessment data and use of assessment data; Considerations in reporting and dissemination; Item development and test design; Technical sampling integrity, efficiency and design; Use of performance standards or achievement levels; and Measurement of test score growth.

The NSC is rated as emerging as a measurement tool for learning assessment at country level, although it is established in terms of the level of development of its security, standardisation, analysis and reporting of the examinations, especially in relation to its role in certification of learning after twelve grades. Despite being an established examination at Grade 12, the NSC is inadequate for measuring learning progress in the whole of the system (in addition to Grade 12) as reflected in its rating on the technical modifications to the assessment quality dimension. By its nature, piloting is impossible, and it also requires confidential and objective development of

[74] It is acknowledged that the technical sub-dimensions and criteria added may have been implicit in the original framework (SABER, 2014), however, they are specified and emphasised here because they relate to specific aspects of the South African experience of technical implementation of assessments, and they are included in the technical sub-dimensions in the rubric since they are relevant to developing countries in particular, and they may be incorporated into the relevant assessment quality sub-dimensions in the original framework.

good quality test items with the correct range of cognitive difficulty and depth across years (DBE, 2016). In addition to monitoring the implementation of the various processes of the NSC examination, monitoring of the examination includes process audits of centres, appointment procedures for markers, invigilator training and support, system readiness to administer the examinations, and monitoring of writing and marking. Risk management of the sites and processes of examination has been the main focus of the standardisation of the examination. School-based assessment processes and systems are also monitored by Umalusi for quality assurance of the certification as they constitute part of the final mark; in some practical subjects, the contribution may be more than 25%. Increasing reporting of quality assurance, monitoring and deviations from standard processes has been a feature of the NSC in recent times. Reduced irregularities, the introduction of training, monitoring of a marking tolerance range for discrepancies in mark allocation per question, the involvement of state security and the police in maintaining process, and test security in the transfer of materials from districts to schools on examination days all show improved attention to security and standardisation (DBE, 2016).

There has been an increase in the credibility and maturity of NSC security processes for standardisation, analysis and reporting of NSC results. However, the DBE notes concerns around test design and development processes (DBE, 2015c). NSC item performance and difficulty rely on panels of marking and moderating specialists rather than on statistically-derived information on item performance. Item-level marker consistency and difficulty has not been standardised empirically between years, and this has been highlighted by Umalusi and the DBE, as a concern (DBE, 2015c) and area of development.

TIMSS and PIRLS are fully established standardised assessments tending with a full range of online documentation and guidelines for standardisation, sampling, analysis and reporting including international versions of the reports which are available online and through the IEA.[75] There are clear protocols for exclusion and inclusion of the participants. Information on schooling context, teacher preparedness and

[75] Available on www.timss.bc.edu

learner background is available. Reporting of mathematics and science scores has been by socioeconomic status, sub-national area, school type, gender, language of the test, cognitive area, domain and question type (Mullis et al, 2011; Performance benchmarks were set to advanced, high, intermediate or low proficiency based on expected levels of curriculum mastery in the learning domains assessed for minimum competency (Reddy et al, 2012). The recently released TIMSS 2015 results indicate an improvement in South Africa's performance in mathematics and science between 2002 and 2015, the largest absolute improvement among the participating countries albeit off a low base (Reddy et al, 2016). PIRLS and TIMSS ratings in this modified framework exceeded that in the original framework due to their technical characteristics which are more developed than the policy- and decision- support applications in South Africa.

A default latent score was the rating for the MLA as it is no longer administered and was only carried out once (DoE, 1999). Documentation other than the MLA country report by the service provider could not be traced. There was no evidence about test confidentiality or security, although independent service providers carried out the work. In addition, apart from the information on data collection training administered by the service providers to officials cascaded through provinces, processes for standardisation were not clear. No information on test and item design, test development or performance standards could be extracted. Basic reporting using classical scoring (percentage correct) was used throughout the report. Neither the background questionnaires nor the instruments used could be examined. MLA was not intended for measuring learning progress and this is reflected in the low ratings allocated to the MLA programme.

The well-known SACMEQ survey was first administered two decades ago as a regional assessment to provide comparative information on learning outcomes in mathematics, language and HIV-AIDS knowledge among Grade 6 learners SACMEQ (Moloi & Chetty, 2010; Ross et al, 2005). Documentation up to the 2000 implementation of the SACMEQ programme is comprehensive, with gaps in technical documentation thereafter especially in respect of analysis and scoring. This situation

does not allow the rating to exceed that of emerging in terms of the technical modifications made to the evaluation matrix used in this study. In recent years, methodology and documentation on test score growth analysis, sampling, analysis and standardisation processes have not been available, with the assumptions over the last decade being that, as the majority of items have remained the same, these have remained the same as in the first two SACMEQ exercises. TB knowledge items were included in 2013, security was highlighted in training and the administration overseen by a national research coordinator based in the DBE. Confidential instruments were used in the surveys and item and test design for 2007 and 2013 were similar, with a small number of items replaced as there was evidence that a few items had been released by researchers without regard to due process. The same Grade 6 target population was sampled. Technical documentation on data processing and analysis in SACMEQ IV, including scaling and standardisation of scores, was not available at the time of writing for country-level checking.[76] It is anticipated by the National Research Coordinator in South Africa that these will be used when they are released to replicate the standardisation and analysis and to satisfy technical requirements for standardisation in analysis.

The majority of modified quality sub-dimensions are in an emerging state in the Systemic Evaluation. However, the systemic evaluation reports did not allow for a rating exceeding emerging for most sub-dimensions of assessment quality in the modified framework whether in relation to bias checks, quality assurance, standardisation or comparability over years.

The NSES is rated as tending to established in terms of technical quality since the assessment tools were primarily designed and refined for the purposes of research rather than for measuring the progress in learning (Taylor, N., 2011). Security and confidentiality in the NSES were achieved through independent administration by service providers. The main item and test-related attributes and sub-components of the NSES are identical to those of the Systemic Evaluation as the instruments used were the same. Standardisation was done in relation to training, checking and

[76] Personal communication with Dr M Moloi and Dr M Chetty, South African National Research Coordinating unit, July and August 2016.

tracking of data collection and fieldwork, analysis and reporting. Test design and development did not use Item Response Theory or Rasch methodology although, after the test, these approaches were used to generate ability scores for each student and difficulty scores for each item on the same scale of achievement. The origin of the scale is set arbitrarily at zero and student and item locations are then distributed across this scale to check and group achievement levels of students at an agreed set of cut-off points, as with the NAEP process. As it was a research project, technical documentation on the NSES is detailed and freely available.

Standardisation has not been perfect in ANA and this, as well as the lack of comparability of ANA results between years, explains its emerging status since ANA tests were developed yearly for use (DoE, 2011a; SAB&T, 2013). The appointment of the service provider for the sample-based ANA has frequently been delayed and the sampling methodology and empirical equivalence for tests have changed over the years. For example, in 2011, the sample excluded small schools with fewer than 25 learners in a grade examined (DoE, 2011a). A population probability sampling approach was used in 2013 (SAB&T Deloitte, 2013). This increased the likelihood of schools being chosen if they had large enrolments. Furthermore, the assessment framework changed between these and reporting in 2013 did not emphasise the verification of marking in the analysis. Despite these challenges, considerable efforts have been made to develop parental feedback reports, and diagnostic reports on weaknesses observed in the ANA (DBE, 2011a and 2013). Documentation on instrument bias or differential item functioning was not available at the time of writing. 50% is the arbitrary cut off for satisfactory achievement although this is not related to the actual levels of achievement or mastery of curriculum among the learners who participated in the assessment. Test score growth between years cannot be used to evaluate or pronounce on the progress in learning as the sample-based ANA tests used are not equivalent or of the same level of difficulty in different years (DBE, 2014a). Altogether, these weaknesses contributed to the emergent rating of sample-based ANA.

Conclusion

The review of systemic learning assessments contained in this chapter indicates that on the whole, an enabling environment exists for learning assessment in South Africa's education system but understanding of the role of different types of assessment is not consistent (Cartwright, 2013). The social and technical validation of systemic learning assessments is not optimal; results are more frequently used at high level for system accountability and reporting purposes. The original framework emphasises the use, dissemination and application of results for policy- and decision-support at country level, while the modified framework supplements this with technical sub-dimensions mainly.

Systemic alignment in Clarke's framework implies a coherence of different curriculum-aligned assessment tools in an integrated assessment system which is well understood by all education practitioners and officials, especially teachers. The NSC, SACMEQ and sample-based ANA have the highest three ratings in systemic alignment learning assessment programmes in South Africa mainly because teachers have been exposed to the NSC and both universal and sample-based ANA and officials are involved in SACMEQ administration. International assessments require more advocacy to improve understanding at the country level in provinces, districts and schools. This will strengthen system alignment of the TIMSS and PIRLS assessments even if the curriculum matching is not exact.

The generally low ratings of the assessment quality sub-dimensions in the original and modified version of Clarke's framework relate to the dearth of technical documentation, weak test design and development, the lack of comparability of the tests between years, and inconsistencies in assessment implementation giving rise to standardisation concerns.

The various systemic assessment programmes provide opportunities for learning lessons about the technical requirements for measuring learning outcomes in South Africa. The international assessments, the NSC examination and even the Systemic Evaluation have technical sub-dimensions with high ratings, and both NSC and sample-based ANA are particularly well aligned to the national curriculum. Provided its technical credibility, comparability and test design are improved, sample-based

ANA provides the basis for the measurement of learning outcomes over time at the country level in South Africa, drawing on best practice in the international assessments and the aspects of the national programmes such as NSES and the NSC in relation to the technical implementation and standardisation protocols for administration. Similarly, PIRLS can be strengthened by improving general knowledge of its use, and improving the further analysis of the African language data in-country after each cycle.

Chapter 5 Conclusion: Analysis of systemic learning assessments in South Africa

This chapter outlines policy and further research implications, based on the findings of this study which together, form the conclusions of this study. The study arose from an acute concern about how educational progress is measured in developing countries in general and in South Africa in particular. In examining the current programmes for measuring learning progress the study concludes with a proposal for strengthening the features and processes around the implementation of a sample-based Annual National Assessment (called Verification-ANA or V-ANA) drawing on past experiences in South Africa and best practice in the literature. The modified sub-dimensions in the modified analytical framework produced in this study are informed by evidence in the literature review, and they provide guidance on the specific technical requirements for establishing a functioning and well-designed systemic assessment within a comprehensive assessment framework in the country. The details in the modified framework help to specify areas for strengthening the implementation of country level assessment programmes which may be useful to developing countries wishing to set up effective system-level assessments to measure learning progress. The policy recommendations are made from the weaknesses identified in the literature review and the analysis and profile developed on the assessment programmes in South Africa. They also necessarily include institutional insight drawn from over two decades of experience in the education system in the country.

Following an analysis of different frameworks for classifying the characteristics and levels of development of education assessment systems in Chapter 2, Clarke's framework was identified as most appropriate for classifying the components required for implementing credible standardised systemic learning assessments in developing countries. Clarke draws on experiences and practices in this category of countries and in industrialised countries in the OECD (Rosenkvist, 2010) as well as on best practice from developed and industrialised countries (NCES, 2015; Cresswell et al, 2015; Lockheed, 2008; Braun and Kanjee, 2006; Ferrer, 2006). Clarke refers to the four types of assessment programmes in an effective educational assessment system

in a country – namely: examinations; classroom-based assessment; international large-scale assessments; and national large-scale assessments which assess national education provision and can be used to measure progress in learning against national learning standards at the country level. The study focused on this last type of assessment which is referred to as systemic assessment.

The evidence for the proposed modifications, based on best practice, is presented in Chapter 2. The analysis presented in Chapter 3, using the modified evaluation matrix, identified strengths and weaknesses of past and existing programmes purporting to measure the progress in learning in South Africa's education system.

A chronological profile of systemic learning assessment follows the analysis of South Africa's systemic assessment programmes in Chapter 4 which also chronicled South Africa's extensive participation in international and country-level assessment programme in narrative form. The concluding chapter includes practical proposals for better assessment of learning and makes policy recommendations of the study findings, as well as proposals for further research.

The contributions of this study to research includes the development of the modified framework for analysing systemic learning assessments (also referred to as national large scale assessments in the literature); the presentation of evidence for the technical dimensions of this modified framework; an updated chronology of learning assessments to 2016; and a chronological profile of systemic learning assessment programmes in which South Africa has participated since 1994. The study also provides evidence for a continuum between examinations and learning assessment in education systems.

The study findings confirm the utility of the measurement of progress in learning, although such measurement needs to take account of the complexity of education systems which involve a variety of people, processes and institutions interacting at different levels. Measurement of learning progress at the country level requires appropriately developed tests, standardised administration, empirical analysis, well

informed reporting and other technical dimensions in addition to a supportive enabling context and high levels of system alignment.

Even critics of educational assessment testing agree that controlled and reasonable assessment, aligned to national priorities, is a necessary pre-condition for assessing the progress in learning in a country (Darling-Hammond and Wentworth, 2010). The draft National Integrated Assessment Framework (Motshekga, 2016 and Mweli, 2016) acknowledges this for South Africa and marks the start of a comprehensive approach to assessment reform, building on the policy intention to develop a world class system of assessments which was articulated in the Department of Basic Education in *Action Plan 2014* and subsequent sector plan (DBE, 2010a and 2015a) as well as the first assessment policies developed in the 1990s (DoE, 1998).

5.1 Policy recommendations

This study has responded to the three research questions posed initially:

i. What is known about the origins, use and utility of country-level learning assessment practices internationally? As noted in the previous section, the term is systemic learning assessments.

ii. Using a modification of an existing analytical framework (Clarke, 2012b) as a point of reference, what are the strengths and weaknesses of the systemic assessment programmes for measuring learning over time in South Africa?

iii. What research conclusions and policy insights will strengthen the measurement of learning in order to better track progress in learning outcomes in the basic education system in South Africa?

The recommendations of this study will have benefits beyond only sample-based systemic learning assessment programmes. Due to their technical nature, some recommendations such as those relating to better technical capacity and investment, social validation as well as improving general understanding of the role and linkages between different kinds of assessment are relatively low cost, but they require a

detailed plan and can be achieved in the short term with little more than a consultative approach and iterative soliciting of perceptions from key stakeholders. Improving documentation, test design, standardisation and comparability of assessments may be achievable in the short to medium term with low to moderate investment and dedicated attention to capacity building. These technical improvements have benefits beyond just the sample-based systemic assessments which are the focus of this study.

The case for credible, technically valid sample-based assessment of learning outcomes as part of a cost effective and comprehensive basket of assessment types in an education assessment system is convincingly made in the literature. Developing countries require the development of the necessary incentives, guidance and support to teachers, parents and officials on assessment practices. Countries also need to establish the necessary monitoring programmes to discourage perverse behaviour, cheating or exclusion of poor performers. In addition, countries need to ensure the development of a value-added basket of indicators of educational performance in the early, intermediate and higher grades of school. This will aid in the development of a credible learning assessment system in the medium to long term.

Assessment reform requires change and process management including broad agreement on the use of data, technical improvements and a shared understanding of the steps and prerequisites for implementation, review and refinement of the system interventions proposed over a period of 10 to 15 years. In their examination of Uganda's assessment and examination system, Allen, Elks, Outhred and Varly (2016) caution that official edict and instructions rarely lead to sustainable change. A compact for change is essential, in tandem with any technical interventions to be made in assessment reform. Teachers need to be at the centre of this compact. The values, attitudes, intentions and perceptions of all players, including teachers, parents, officials and learners involved in assessment (and curriculum) reform, need to be understood and concerns acted on and incorporated into the change process to institutionalise the reform.

The recommendations emerging from this study are listed below:

i. **Assessment system with variety of tools**. A better understanding of the place, role and limitations of sample-based assessments is necessary at all levels in the education system in South Africa. System level score growth from a sample-based assessment is indicative of country-level progress in learning over time and will be limited to reporting on groups of learners, with limited information provided about general weaknesses in learning against national curriculum standards. The sample, analysis and reporting should reflect the limitations of the system-level sample-based information provided for learner level analysis, but should explicitly show the links with other (non-sample based) assessments in the education system and thee roles of both for monitoring and accountability at the appropriate level. It is important, however, that the validity of the assessment programme is not overshadowed by reliability concerns, no matter what the assessment functions as in the system.

ii. **Strengthened capacity.** Capacity building programme on using advanced techniques in data analysis, psychometric analysis of items and the calculation of standardised and classical test scores need to be carried out on a sustained basis. This understanding needs to be expanded and more clearly understood by parents, learners, teachers, officials and policy makers. A sustained learning and training programme should be formulated to enable skills development at emerging, intermediate and advanced levels of measuring learning achievement including an emphasis on developing documentation, guidelines and specifications in items and test development and implementation. The unique political support of the SACMEQ programme at regional level may be used to facilitate the development of regional capacity, provided investments in technical capacity and resourcing can be used to urgently remediate the weaknesses identified in SACMEQ. Alternatively, other agencies working in UNESCO International Institute for Education Planning (IIEP) and other UN agencies involved in measuring and monitoring learning and development outcomes should also consider regional efforts in support of improved learning progress

measurement. It may be possible, due to the nature of the spill-over effects of improving the measurement of learning outcomes, to mobilise regional support and technical assistance to audit the technical levels of development of education assessment systems in the region and to report on best practice in assessment practice and technical implementation within the region and in individual countries especially with respect to basic reading, literacy and numeracy competencies.

iii. **Diverse monitoring tools.** The systemic monitoring of progress in learning outcomes through sample-based assessments should happen concurrently with other forms of monitoring at classroom and school level which would enable learner level monitoring information to be collected and reported in an education system.

iv. **Further research and analysis.** Research and analysis is needed to support assessment and examination design, implementation, and perceptions in the sector with adequate attention and resourcing directed to such analysis and research. This needs to be well resourced, and guided by the Department of Basic Education in partnership with researchers in higher education institutions. In the short term, regional and national audits of learning assessment instruments and programmes used for different functions by country should be carried out with a view to understanding assessment relationships, intent and methodology at national and sub-national level building on the work of Benavot and Koselici, 2015 and Clarke (2012b) including the modifications recommended and used in this study. At country level, a repository of assessment instruments should be developed confidentially by grade, subject and construct. A research agenda to support assessment reform will have benefits for deepening assessment use and utility in school and beyond, since post-school articulation will be expected to include assessment articulation and alignment as well as instruction and learning programme delivery in class. The support of the Department of Science and Technology and research agencies such as the National Research Foundation in such studies to benefit schooling and post-

schooling outcomes in South Africa is particularly important. More postgraduate research should be guided to focus on issues of learning outcomes measurement through engagement with the National Research Foundation and agencies involved in knowledge creation and development.

v. **Social validation.** Social validation must receive attention – informed by sufficient evidence of technical information on assessment tests, processes and requirements. The experiences of other developing countries such as Uruguay and Brazil (Ferrer, 2006 and de Castro, 2012) indicate that effective reform takes time and requires well-evaluated social validation processes in addition to credible rational planning and monitoring mechanisms.

vi. **Planning of assessment system development and reform.** Assessment reform in the schooling system requires coherence and elaboration with a theory of change for the outcomes of each of the sub-systems so that the links between different types of assessment are clearly articulated and planned for. An audit of assessment practice is proposed to supplement this and to provide in-depth understanding of education assessment policy in the country's provinces, districts and schools. It is anticipated that the School Monitoring Survey, a sample-based survey, will capture some of this information if it is implemented successfully in 2017 following from the 2011 School Monitoring Survey.

vii. **Item, test development and design.** Expertise in the area of item and test development has been consistently identified as a weakness in all types of assessment in the country. National, regional and international capacity development partnerships are needed to address this priority area and require curriculum, assessment, psychometric and language specialists to work together. The goal should be the enhancement of test development and other technical skills in the country among curriculum, assessment and education evaluation practitioners and researchers. Assessment tests and items need to be developed for diagnostic, formative and summative assessment functions in languages, and life skills for the early grades, especially in African Languages. These tests and teacher development material on the relevant assessment tests need to be

quality assured, psychometrically and cognitively appropriate for the constructs being assessed, and modified for the South African context and normed for African Languages. Materials for teacher support are required to complement such tests in order to assist teachers to understand and remediate weaknesses and barriers to learning progression in the classroom (Heritage, 2008).

viii. **Standardisation and bias testing improvements**. Improving standardisation and the integrity of the administration of different types of assessments and examinations is an urgent priority. The administration of international assessments provide a benchmark for standardisation and these may be used to improve standardisation of administration processes especially in relation to protocols for the use of officials and independent service providers, development of specifications for standardised administration, and checking for bias within the implementation processes. The integrity of examinations administration and standardised assessment administration requires monitoring in all countries. Incidences of cheating and manipulation of test scores through exposure of confidential test instruments is a feature of education systems in developing and industrialised countries alike. To secure gradual improvements in standardisation of the implementation of learning assessments, standardisation of assessment testing may focus initially on issues of operations, basic administration and process fidelity and then assessment content, analysis and reporting in that order.

ix. **Item banking.** Country-level investments in item banking[77] needs to be rationalised and quality assured to ensure that items in item banks address the full range and diversity of learning domains, learner ability and cognitive difficulty. Summative assessment items tend to be more focused and selective in terms of complexity and curriculum domains covered. Items intended for diagnosis should explore deeper learning deficits, while other items may fulfil all more than one purpose. For developing countries, clear specification and quality assurance of all banked items must be upheld or the utility of banked items will

[77] An item bank is a repository of test items that belong to a testing program, as well as all information pertaining to those items.

be limited. South Africa's Teacher Assessment Resource for Monitoring and Improving Instruction (TARMII) project has been in existence since 2002 but still needs to be independently evaluated and all items specified using psychometric and empirical means to ensure cognitive and learning performance for each item is understood.

x. **Special needs learners and assessment.** A special research effort should be embarked on to investigate the assessment of learners with special needs including those with disabilities and learning deficits. This should focus on the adaptation, modification and application of assessment instruments and tools to enable more authentic assessment of these learners, starting with conceptual development deficits which are barriers to future learning for many of the learners especially in the areas of literacy and numeracy. These will build on the efforts of the Foundation phase Systemic Evaluation which were carried out in special schools a decade and a half ago.

xi. **Participation in Programme for International Student Assessment for Development (PISA-D).** Assessing the practical skills of 15 year-olds through PISA for Development should be considered in the short term as this would be an opportunity to monitor the implementation of the technical and vocational curriculum which has recently been expanded throughout the country. South Africa would do well in benchmarking performance through continued participation in TIMSS, PIRLS, SACMEQ and PISA-D. PISA-D is geared towards practical problem-solving, critical and other high-demand 21st century skills which position learners well in the global knowledge economy. Household-based surveys such as the Annual Status of Education Report (ASER) in India or Uwezo in Kenya are not appropriate for countries with a high participation and enrolment rate such as South Africa. Regional groupings can assist in this respect as they can introduce economies of scale into such technical undertakings. It is understood from the National Country Coordinator, that South Africa has begun this process to assist with comparisons of country performance across different assessments.

xii. **Technical and administrative oversight should be separated and clearly linked.** An overarching recommendation of this study is that the planning, management, governance and oversight of assessment reform must focus on monitoring learning outcomes over time, and they must be distinct and not conflated. In the case of ANA, according to the official responsible for assessment implementation, an advisory committee was set up with wide ranging responsibilities for advice but the configuration of technical specialists and officials and administrators could not provide sufficient capacity to avoid the weaknesses identified in the analysis carried out in this study. A team of specialists at regional level may be made available to consultation with countries in an effort to improve this oversight.

xiii. **Assessment reform plan**. Finally, a medium- to long-range basic education sector assessment road map and plan should be developed with roles, responsibilities and resourcing explicitly stated. This plan must incorporate as priorities the different aspects of the assessment and evaluation of classroom performance, teacher performance, school performance and sector performance. This plan should include the change management and technical dimensions identified in this study in relation to enabling context, system alignment and assessment quality. Such a sector plan requires a diagnostic analysis and audit of the use of, and linkages between, assessment instruments, processes and information at different levels in the system. A theory of change for assessment at different levels will assist in the development of this sector plan, which should include the technical and non-technical tools, diversity of mechanisms and some of the components and dimensions identified in the literature and in the modified framework used in this study.

5.2 Opportunities for further research

Based on the main findings and gaps identified during the study, the following suggestions regarding further research are presented in this section. The literature confirms that education reforms involving the use of learning assessments enable policy makers in education systems to monitor trends in learning achievement.

For example, such research may focus on methodological issues of standardisation and bias checking in assessment data, better and more contextualised test design in the foundation phase of school, testing of learners with special needs, investigating the effects of assessment data and information on practice and behaviour in favour of, and away from quality-seeking activities. Implementation research is also needed into how to progressively reduce bias in assessment testing over time, and urgent work needs to be done on developing school performance indicators which are stable and reflective of quality of educational performance and which include learning outcomes and assessment information adjusted for context and learner population served.

Given that many developing country education systems exhibit low levels of learning achievement, the lack of detail on the mechanisms and methods for translating the findings of learning assessment into instructional improvement through institutionalised teacher development and instructional strengthening initiatives should form the basis for additional research.[78]

More research is required into the effects of translation of test items on the test scores and performance of the test taking population in different languages, especially in African languages. Differential Item Functioning of items translated into different African languages has been observed with PIRLS 2006 data and this may have introduced bias in the assessment data for different languages. Many assessments use straight translations which are less expensive but also less culturally contextualised. Although versioning is more labour intensive and costly, it is particularly important in the South African context which has language diversity and a lack of reading materials and literacy scaffolding outside of the main language groupings.

[78] In many developing countries, teachers' assessment skills are not optimal and classroom practice leaves a lot to be desired. Persistent challenges in schools disrupt the coherence between teaching, the curriculum and assessment. Mcintosh (1994) in a wide ranging study drawing on international experiences, laments the poor preparation of teachers skills in the area of assessment as necessary and critical and makes recommendations for better, more institutionalised teacher capacities for undertaking assessment, using and applying assessment results in efforts to improve equity in learning achievement.

Finally, in developing item banks in different contexts, test item performance and dispersion in different contexts in similar contexts should be used to compare and document how test items perform in different contexts. At regional level, this would be most useful in a group of countries aided by regional co-operation between development partners including UNESCO, UNICEF, and regional assessment programmes such as *Programme d'analyse des systèmes éducatifs de la CONFEMEN* (PASEC)[79] and SACMEQ which may consider this as a project. Drawing on TIMSS, SACMEQ, PASEC expertise and item development expertise with an explicit focus on, for example, common learning constructs and items related to the particular numeracy and literacy weaknesses affecting countries on the continent, such a project would assist development co-operation between countries and even within the region in support of global education development goals.

Further research should be targeted at ensuring that the voices of all involved in assessment in the education system are heard in relation to their past and current perceptions and actions. Such perceptions may be canvassed in relation to the burden, depth, nature and coverage of different assessment programmes including those not explicitly covered in this study, namely: classroom; school; and, teacher performance assessment and evaluation systems.

Provincial assessment systems need to be understood at the country level as these systems may be useful in the technical implementation of programmes for measuring progress in learning outcomes, and for monitoring education quality in the country's education system.

5.3 Limitations

This study used evidence to examine the usefulness of the framework proposed by Clarke for assessing educational assessments. The evidence was drawn from technical documents and from literature on South African implementation of such programmes where relevant. The study might have been strengthened by qualitative interviews with key informants to reflect on developments and experiences of assessment in the education sector in South Africa, however, in the absence of such

79 PASEC is carried out in nine francophone mainly West African countries since 1995.

interviews, the author was able to draw on her extensive experience in the education sector to confirm some of the findings in the literature. The ratings allocated, in the modified evaluation matrix, to the systemic assessment programmes in South Africa were tested against the literature and the institutional memory of the development of the assessment programmes. Undocumented information on the latter as well as interviews with a number of individuals, were included. The latter, as personal communication, albeit in non-systematic interviews.

The weightings allocated in the evaluation matrix may not have been ideal for the South African context and typically should be subject to consultation with policy makers in terms of relative weighting of the three dimensions of assessment in order to fine-tune the ratings since these are always subject to contestation. This is a possible area of research for the future as part of the deepening of assessment reform in the country, although Clarke's method assumes equal weighting to all three dimensions.

The dearth of research and information on teachers' and parents' perceptions of assessment practice, especially in recent years in developing countries, was a significant challenge to the study. Research has typically been focused on formative assessment at classroom level. The lengthy literature review in this study assisted in formulating the evidence for the framework used in the study and enabled the analysis to be completed.

Digitally available literature and technical documentation in English from developed and developing countries was reviewed, although the technical documentation on standardised international assessment implementation was not widely available. Where it was available, it was limited in relation to the specific assessment dimensions of quality discussed in Chapters 3 and 4. There is lack of documentation in the public domain about technical methods and processes used in education assessment globally with the notable exception of some countries in the OECD, Europe, Latin American and the United States. This may be the result of caution on the part of countries still in the process of assessment reform. The experiences of

many developing countries are largely absent from the literature, a state of affairs that will probably change as the new Sustainable Development Goals require better measurement of learning outcomes and education system quality in future.

5.4 Conclusion

The global interest in the assessment of learning outcomes in developing countries arises from the increasing importance attached to these outcomes. The judicious use of systemic learning assessment tests can benefit education system performance measurement by providing information on knowledge acquired in schools which may then be used for accountability and monitoring purposes at the country level. The latter purpose forms the main focus of the study.

This study contributes in three ways to the literature on learning assessment. Firstly, it examines what is known about the technical features of effective systemic learning assessment programmes internationally. The study focused on sample-based assessments used for measuring learning progress at the country level as these are most cost efficient for country level measurement of learning outcomes and their administration can easily be kept secure and confidential. Secondly, it examines the strengths and weaknesses of past systemic learning assessment programmes in South Africa's schooling system, and provides evidence for the relevance of a modified framework which was used to analyse the assessments. The resulting analysis provides the basis for developing recommendations on how existing assessment programmes may be strengthened in order to secure technical improvements in the measurement of learning progress.

The findings of this study suggest that a more coherent approach to assessment system development is required, and proposes refinements and modifications to an existing sample-based assessment programme. The technical recommendations made in the study can be applied beyond just sample-based systemic assessments,

and they provide a basis for strengthening the measurement of the progress in learning outcomes at the country level. The recommendations may be applied to other developing countries as they include proposals for improving the capacity and fidelity of standardised learning assessment implementation at the country level. Specific recommendations include improving test design for comparability across years, enhancing curriculum-aligned item development, strengthening governance and integrated planning of educational assessment reform, and deepening the understanding of the role of, and linkages between, different assessment programmes in the education system and more broadly. The recommendations focus on adapting and refining an existing sample-based assessment programme using the lessons learned from the experiences of the past two decades in the country. The caveat is that the assessment system must be focused on what is relevant and valid in terms of what is measured, rather than what is efficient and convenient to measure reliably.

Chapter 6 References

Alberta Education. (2013). How the Accountability Pillar Works. Retrieved from http://education.alberta.ca/admin/funding/accountability/works.aspx.

Allen, R., Elks, P., Outhred, R., & Varly, P. (2016). Uganda's assessment system: a road-map for enhancing assessment in education. Final assessment report prepared for the Health and Education Advice and Resource Team (HEART). September 2016.

ANC. (1994). Policy Framework for Education and Training. Chapter 13. African National Congress, South Africa. http://www.cepd.org.za/files/pictures/Policy%20Framework%20For%20Education%20and%20Training%20ANC.PDF

Bandeira de Mello, V., Blankenship, C., & McLaughlin, D. (2009). Mapping State Proficiency Standards onto NAEP Scales: 2005-2007. Research and Development Report. NCES 2010-456. National Center for Education Statistics.

Bandeira de Mello, V., Bohrnstedt, G., Blankenship, C., & Sherman, D. (2015). Mapping State Proficiency Standards onto NAEP Scales: Results from the 2013 NAEP Reading and Mathematics Assessments. NCES 2015-046. National Center for Education Statistics.

Barber, M., & Mourshed, M. (2007). How the world's best-performing schools systems come out on top. McKinsey & Company.

Barton, P. E. (1999). Too Much Testing of the Wrong Kind; Too Little of the Right Kind in K-12 Education. A Policy Information Perspective.

Benavot, A., & Köseleci, N. (2015). Paper commissioned for the EFA Global Monitoring Report 2015, Education for All 2000-2015: achievements and challenges Seeking Quality in Education: The Growth of National Learning Assessments, 1990-2013. ED/EFA/MRT/2015/PI/53.

Best, M., Knight, P., Lietz, P., Lockwood, C., Nugroho, D., & Tobin, M. (2013). The impact of national and international assessment programmes on education policy, particularly policies regarding resource allocation and teaching and learning practices

in developing countries. Final report. London: EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Bishop, J. H. (1997). The effect of national standards and curriculum-based exams on achievement. The American Economic Review, 87(2), 260-264.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. Assessment in Education: principles, policy & practice, 5(1), 7-74.

Black, P., & Wiliam, D. (2007). Large-scale assessment systems: Design principles drawn from international comparisons 1. Measurement, 5(1), 1-53.

Boissiere, M. (2004). Determinants of primary education outcomes in developing countries. Determinants of Primary Education Outcomes in Developing Countries.

Bovens, M. (2005). 8.1 The concept of public accountability. *The Oxford handbook of public management*, 182.

Braun, H., & Kanjee, A. (2006). Using assessment to improve education in developing nations. Improving education through assessment, innovation, and evaluation, 1-46.

Braun, H., Kanjee, A., Bettinger, E., & Kremer, M. (2006). Improving education through assessment, innovation, and evaluation. Cambridge, MA: American Academy of Arts and Sciences.

Brennan, R. L. (2004). Revolutions and evolutions in current educational testing. Center for Advanced Studies in Measurement and Assessment: CASM Research Report, 6.

Brookhart, S. M. (2013). The public understanding of assessment in educational reform in the United States. Oxford Review of Education, 39(1), 52-71.

Bruns, B., Evans, D., & Luque, J. (2012). Achieving world-class education in Brazil. World Bank Publications.

Bruns, B., Filmer, D., & Patrinos, H. A. (2011). Making schools work: New evidence on accountability reforms. World Bank Publications.

Burgess, S., Wilson, D., and Worth, J. (2013) A natural experiment in school accountability: the impact of school performance information on pupil progress and sorting. Journal of Public Economics 106, 57 – 67.

Carnoy, M., Chisholm, L., Addy, N., Arends, F., Baloyi, H., Irving, M., …. Sorto, A. (2011). The Process of Learning in South Africa: The Quality of Mathematics Teaching in North West Province. HSRC Press.

Carnoy, M., Chisholm, L., & Baloyi, H. (2008). Towards understanding student academic performance in South Africa: a pilot study of grade 6 mathematics lessons in Gauteng province. HSRC, South Africa.

Carnoy, M., Chisholm, L., & Chilisa, B. (2012). The low achievement trap: Comparing schooling in Botswana and South Africa. HSRC Press.

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. Educational evaluation and policy analysis, 24(4), 305-331.

Cartwright, F. (2013). Review of the national ANA programme, South Africa. World Bank 2013.

Chakwera, E., Khembo, D., & Sireci, S. G. (2004). High-Stakes Testing in the Warm Heart of Africa: The Challenges and Successes of the Malawi National Examinations Board. education policy analysis archives, 12(29), no29.

Chinapah, V., H'ddigui, E. M., Kanjee, A., Falayajo, W., Fomba, C. O., Hamissou, O., ... & Byomugisha, A. (2000). With Africa for Africa. Towards Quality Education for All. 1999 MLA Project. Human Sciences Research Council, Private Bag X41, Pretoria, South Africa, 0001; Web site: http://www. hsrc. ac. za.

Clarke, M. (2012a). Measuring learning: how effective student assessment systems can help achieve learning for all (No. 10058). The World Bank.

Clarke, M. (2012b). What matters most for student assessment systems: A framework paper. SABER–Student Assessment Working Paper, 1.

Conley, D. (2015). A New Era for Educational Assessment. education policy analysis archives, 23(8), n8

COSATU. (2011). Congress of South African Trade Unions. SADTU welcomes the state of the nation address. COSATU Today. COSATU Press Statements. Retrieved February 10, 2017 from: http://www.cosatu.org.za/show.php?ID=4484

Cresswell, J., Schwantner, U., & Waters, C. (2015). A Review of International Large-Scale Assessments in Education: Assessing Component Skills and Collecting Contextual Data. PISA for Development. OECD Publishing. 2, rue Andre Pascal, F-75775 Paris Cedex 16, France.

Crouch, L., & Mabogoane, T. (1998). When the residuals matter more than the coefficients: An educational perspective. Journal for studies in economics and econometrics, 22, 1-14.

DA Press Statement. (2014). SADTU is right: Annual National Assessments aren't working. Retrieved February 2017 from http://www.da.org.za/2014/10/sadtu-right-annual-national-assessments-arent-working/

Darling-Hammond, L., & Wentworth, L. (2010). Benchmarking learning systems: Student performance assessment in international context. Stanford Center for Opportunity Policy in Education, Stanford University, CA.

Darling-Hammond, L., Wilhoit, G., & Pittenger, L. (2014). Accountability for College and Career Readiness: Developing a New Paradigm. education policy analysis archives, 22(86), n86.

Dee, T. S., & Jacob, B. (2009). The impact of No Child Left Behind on student achievement. Journal of Policy Analysis and management, 30(3), 418-446.

de Castro, M. H. G. (2012). Developing the enabling context for student assessment in Brazil. World Bank.

Deming, D. J., Cohodes, S., Jennings, J., & Jencks, C. (2013). School accountability, post-secondary attainment and earnings. Review of Economics and Statistics, (0).

Department of Basic Education. (2010). Action Plan to 2014: Towards the Realization of Schooling 2025. Pretoria: Department of Basic Education.

Department of Basic Education. (2010b). Curriculum News. Pretoria

Department of Basic Education. (2011a). Report on the Annual National Assessments of 2011. Pretoria: Department of Basic Education.
Department of Basic Education. (2011b). Annual National Assessments 2011: A guideline for the interpretation and use of ANA results.

Department of Basic Education. (2012a). General Education Quality Analysis & Diagnosis Framework (GEQAF). Pretoria

Department of Basic Education. (2012b). Report on the Annual National Assessments 2012: Grades 1 to 6 & 9. Pretoria: Department of Basic Education.

Department of Basic Education. (2012c). Public expenditure analysis report for the Department of Basic Education by Oxford Policy Management and Research on Socio-Economic Policy (ReSEP). Unpublished.

Department of Basic Education. (2013). Report on the Annual National Assessments 2013: Grades 1 to 6 & 9. Pretoria: Department of Basic Education.

Department of Basic Education. (2014a). Report on the Annual National Assessments 2014: Grades 1 to 6 & 9. Pretoria: Department of Basic Education.

Department of Basic Education. (2014b). National Senior Certificate Examination Report. Pretoria: Department of Basic Education.

Department of Basic Education. (2015c). National Senior Certificate Examination Report. Pretoria: Department of Basic Education.

Department of Basic Education. (2015a). Action Plan to 2019: Towards the Realisation of Schooling 2030.

Department of Basic Education. (2015b). Review of implementation of institutional mandates (articulated through plans and reports) for education entities in support of sector planning, monitoring and evaluation. Unpublished document.

Department of Basic Education, 2016. 2016 National Senior Certificate Examination Report. Pretoria: Department of Basic Education.

Department of Basic Education, 2017. Annual Report of the Department of Basic Education. Pretoria: Department of Basic Education.

Department of Education. (1995). White Paper on Education and Training. Notice 196 of 1995. South Africa.

Department of Education. (1996). National Education Policy Act (NEPA) No 27 of 1996, 24 April. Pretoria: Government Printers.

Department of Education. (1996). South African Schools Act (SASA). Government Gazette No. 84 of 1996. Pretoria: Government Printer

Department of Education. (1998a). Assessment Policy for General Education and Training. Pretoria: Government Printers.

Department of Education. (1998b). South African National Norms and Standards for School Funding. Pretoria: Government Printer

Department of Education. (1999). MLA report. Pretoria, South Africa. Supported by UNICEF and UNESCO. Research Institute for Education Planning.

Department of Education. (2000). Norms and Standards for Educators. Government Gazette Vol 415, No 20844 of 4 February 2000. Pretoria: Government Printers.

Department of Education. (2001). National Policy on Whole-School Evaluation. Government Gazette Vol.433, No. 22512 of July 2001, Pretoria

Department of Education (2002) Overview and Chapter 6: Learner Assessment in the Learning Area Statements, Revised National Curriculum Statement Grades R–9. Pretoria: Government Printers

Department of Education. (2003a). Costs, financing and resourcing of schooling in South Africa, Ministerial review report to Minister of Basic Education. Pretoria.

Department of Education. (2003b). Systemic Evaluation: Foundation Phase (Mainstream). Pretoria: Department of Education.

Department of Education. (2003c). National Report on Systemic Evaluation: Foundation Phase (Learners with Disabilities in Special Schools). Pretoria: Department of Education.

Department of Education. (2004). Strategic Plan 2004 – 2006. Pretoria: Department of Education.

Department of Education (2005). Grade 6 Systemic Evaluation: Intermediate. National Report. Pretoria. Available from: http://www.hsrc.ac.za/research/output/outputDocuments/3580_Grade6National.pdf

Department of Education. (2007a). National policy on assessment and qualifications for schools in the general education and training band. Pretoria: Department of Education.

Department of Education. (2008a). Grade 3 Systemic Evaluation 2007 Leaflet. Pretoria: Department of Education.

Department of Education. (2009). Report of the Task Team for the Review of the Implementation of the National Curriculum Statement (NCS), Final Report, October 2009. Pretoria.

Di Carlo. M, August 2013. NAEP and Public Investment in Knowledge, by Matthew Di Carlo -- August 8, 2013. [Web log downloaded 7 July 2016]. Retrieved from: http://www.shankerinstitute.org/blog/naep-and-public-investment-knowledge

Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. Practical Assessment, Research & Evaluation, 14(7), 1-11.

Dunne, T., Long, C., Craig, T., & Venter, E. (2012). Meeting the requirements of both classroom-based and systemic assessment of mathematics proficiency: The potential of Rasch measurement theory. *Pythagoras*, *33*(3), Art. #19, 16 pages. http://dx.doi.org/10.4102/pythagoras.v33i3.19

Education International. (2013). Key messages from PISA 2012. An analysis and briefing from Education International. [Press release] Delivered 2 December 2013 by General Secretary, Fred van Leeuwen. Retrieved from [http://download.ei-ie.org/Docs/WebDepot/Circular_PISA2012_E.pdf] Downloaded 8 February 2017.

Equal education. (2011). Equal Education responds to release of ANA Results. Press Statement. 30 June. Retrieved from: [http://www.equaleducation.org.za/article/ee-responds-to-release-of-annual-national-assessment-ana-results]15 August 2016

Ferrer, J. G. (2006). Educational assessment systems in Latin America: Current practice and future challenges. PREAL.

Figlio, D., Ladd, H., (2007). School accountability and student achievement. In: Ladd, H., Fiske, E. (Eds.), Handbook of Research on Education Finance and Policy. Routledge.

Filmer, D., Hasan, A., & Pritchett, L. (2006). A millennium learning goal: Measuring real progress in education. Center for Global Development Working Paper, (97).

Fleisch, B. (2008). Primary education in crisis: Why South African schoolchildren underachieve in reading and mathematics. Juta and Company Ltd.

Foy, P., Galia, J., & Li, I. (2007). Scaling the data from the TIMSS 2007 mathematics and science assessments. TIMSS, 225-279.

Foy, P., Martin, M. O., & Mullis, I. (2010). The limits of measurement: problems with estimating reading achievement in PIRLS 2006 for low performing countries.

Gordon, A. (2009). Restructuring teacher education. *Issues in Education Policy*, Issue Number *6*.Centre for Education Policy Development. Johannesburg.

Gravett and Henning (2014). Presentation by the University of Johannesburg, School of Education on the occasion of the release of the ANA results in 2014. University of Johannesburg. Unpublished.

Greaney, V., & Kellaghan, T. (Eds.). (2008). Assessing national achievement levels in education (Vol. 1). World Bank Publications.

Grek, S 2009, 'Governing by Numbers: The PISA 'Effect' in Europe' Journal of Education Policy, vol 24, no. 1, pp. 23-37. DOI: 10.1080/02680930802412669.

Gustafsson, M. and Moloi, Q. (2011). Thermometer, pill, placebo or virus? The ongoing debate on when to use standardised assessments and how. Unpublished document.

Gustafsson. M. (2014). Unpublished document. Curriculum change in South Africa over the years

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. American economic review, 1184-1208.

Hamilton, L. S., Stecher, B. M., & Klein, S. P. (2002). Making sense of test-based accountability in education. Rand Corporation. Santa Monica, CA.

Hanushek, E. A. (2003). The failure of input-based schooling policies. The economic journal, 113(485), F64-F98.

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? Journal of policy analysis and management, 24(2), 297-327.

Hanushek, E. A., & Wöessmann, L. (2007). The role of education quality for economic growth. World Bank Policy Research Working Paper, (4122).

Harlen, W. (2005). Teachers' summative practices and assessment for learning–tensions and synergies. Curriculum Journal, 16(2), 207-223.

Heritage, M. (2008). Supporting instruction and formative assessment. National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Paper prepared for the Formative Assessment for Teachers and Students (FAST). State Collaborative on Assessment and Student Standards (SCASS) of the Council of Chief State School Officers, Washington, DC

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2(3), 172-177.

Hoadley, U. (2012). What do we know about teaching and learning in South African primary schools? Education as Change, 16(2), 187-202.

Hoadley, U., & Muller, J., (2016). Visibility and differentiation: Systemic testing in a developing country context. The Curriculum Journal, 27, 2, 272-290.

Howie, S. (2003). Conditions of schooling in South Africa and the effects on mathematics achievement. Studies in Educational Evaluation, 29: 227-241.

Howie, S., Barry, D., & de Kock, H. (2011). Review of the Annual National Assessments for 2011: For the National Basic Education. Centre for Evaluation & Assessment, University of Pretoria, Pretoria.

Howie, S., & Hughes, C., 2000. “South Africa.” In The Impact of TIMSS on the Teaching and Learning of Mathematics and Science, ed. D. Robitaille, A. Beaton, and T. Plomp, 139–45. Vancouver, BC: Pacific Educational Press.

Howie, S. & Pietersen, J.J. (2001). Mathematics literacy of final year students: South African realities. Studies in Educational Evaluation, 27: 7-25.

Howie, S., & Van Staden, S. (2012). South African Children’s Reading Literacy Achievement–PIRLS and prePIRLS 2011 Summary of the Key results (Media briefing). *Pretoria: Centre for Evaluation and* Assessment.

Howie, S., Venter, E., Van Staden, S., Zimmerman, L., Long, C., Du Toit, C., et al. (2008). PIRLS 2006 Summary Report: South African Children's Reading Literacy Achievement. Pretoria: Center for Evaluation and Assessment. University of Pretoria

Hoxby, C. (2001). Testing is About Openness and Openness Works. Hoover Daily Report, July, 30.

Hoxby, C. M. (2002). The cost of accountability (No. w8855). National Bureau of Economic Research.

Huitt, W., Hummel, J., & Kaeck, D. (2001). Assessment, measurement, evaluation, and research. Educational Psychology Interactive. Valdosta, GA: Valdosta State University. Retrieved [1 October 2016], from http://www.edpsycinteractive.org/topics/intro/sciknow.html

Huitt, W. (2003, June). Assessment, measurement, evaluation, and research: Types of studies in scientific research. Educational Psychology Interactive. Valdosta, GA: Valdosta State University. Retrieved [1 October 2016], from http://www.edpsycinteractive.org/topics/intro/research.html

Huitt, W. (2007). Assessment, measurement, and evaluation: Overview. Educational Psychology Interactive. Valdosta, GA: Valdosta State University. Retrieved [1 October 2016], from http://www.edpsycinteractive.org/topics/measeval/msevlovr.html

Huitt, W. (2011). Why study educational psychology? Educational Psychology Interactive. Valdosta, GA: Valdosta State University. Retrieved [1 October 2016], from http://www.edpsycinteractive.org/topics/intro/whyedpsy.html

Hull, J. (2007). Measuring student growth: A guide to informed decision making. *Centre for Public Education. Retrieved on March*, *3*, 2009.

HSRC. (1996). Schools Register of Needs. Human Sciences Research Council. 1996

Hungi, N., Makuwa, D., Ross, K., Saito, M., Dolata, S., Van Capelle, F., & Vellien, J. (2011). SACMEQ III project results: Levels and trends in school resources among SACMEQ school systems. *Paris: SACMEQ*.

Jacob, B. (2005). Accountability, incentives and behavior: Evidence from school reform in Chicago. *Journal of public Economics*, *89*(5-6), 761-796.

Jacob, B. A., & Levitt, S. D. (2003). *Rotten apples: An investigation of the prevalence and predictors of teacher cheating* (No. w9413). National Bureau of Economic Research.

Jerrim, J. (2013). The reliability of trends over time in international education test scores: Is the performance of England's secondary school pupils really in relative decline. *Journal of Social Policy*, *42*(2), 259-279.

Joncas, M., & Foy, P. (2011). Sample design in TIMSS and PIRLS. *Methods and Procedures in TIMSS and PIRLS*.

Kamens, D. H., & Benavot, A. (2011). National, regional and international learning assessments: Trends among developing countries, 1960–2009.*Globalisation, societies and education*, *9*(2), 285-300.

Kamens, D. H., & McNeely, C. L. (2010). Globalization and the growth of international educational testing and national assessment. *Comparative Education Review*, *54*(1), 5-25.

Keeves, J. P. (1994). *National examinations: design, procedures and reporting* (Vol. 50). UNESCO.

Kennedy, M. M. (1990). *A survey of recent literature on teachers' subject matter knowledge* (Vol. 90, No. 3). East Lansing, MI: National Center for Research on Teacher Education.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of human resources*, 752-777.

Koretz, D. (2003). Using multiple measures to address perverse incentives and score inflation. *Educational measurement: issues and Practice*, *22*(2), 18-26.

Koretz, D. (2008). Test-based educational accountability. Research evidence and implications. *Zeitschrift für Pädagogik*, *54*(6), 777-790.

Koretz, D. M., & Barron, S. I. (1998). *The Validity of Gains in Scores on the Kentucky Instructional Results Information System (KIRIS)*.

Koretz, D. M., & Hamilton, L. S. (2003). *Teachers' responses to high-stakes testing and the validity of gains: A pilot study*. Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California, Los Angeles.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika*, *79*(2), 210-231.

La Belle, T. J. (1982). Formal, non-formal and informal education: A holistic perspective on lifelong learning. International review of education, 28(2), 159-175.

Levitt, R., Janta, B., & Wegrich, K. (2008). Accountability of teachers: Literature review. Rand Corporation.

Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment*, *7*(1), 29-38

Linn, R. L. (2005). Issues in the Design of Accountability Systems. CSE Technical Report 650. *National Centre for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Lockheed, M. (2008). Measuring Progress with Tests of Learning: Pros and Cons for Cash on Delivery Aid in Education. *Center for Global Development Working Paper*, (147).

Lockheed, M. E., & Verspoor, A. M. (1991). *Improving primary education in developing countries*. Oxford University Press for World Bank.

Loveless, T., Costrell, R. M., & Cuban, L. (2005). Test-based accountability: The promise and the perils. *Brookings papers on education policy*, (8), 7-45.

Lubisi, R. C., & Murphy, R. J. (2002). Assessment in South African schools. Assessment in Education: Principles, Policy & Practice, 9(2), 255-268.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), Eighty-seventh year book of the National Society for the Study of Education: Part 1. Critical issues in curriculum (pp. 83–121). Chicago: University of Chicago Press. 1988

Makuwa, D. K., & Maarse, J. (2013). The Impact of Large-Scale International Assessments: a case study of how the Ministry of Education in Namibia used SACMEQ assessments to improve learning outcomes. *Research in Comparative and International Education*, *8*(3), 349-358.

Mbanjwa, X., & A Kassiem, A. *SA doesn't need global maths test* – Pandor, The Star, 24 April 2007. Retrieved August 2016 from [https://www.iol.co.za/news/south-africa/sa-doesnt-need-global-maths-test---pandor-349943].

McEwan, P. J. (2015). Improving Learning in Primary Schools of Developing Countries a Meta-Analysis of Randomized Experiments. *Review of Educational Research*, *85*(3), 353-394.

Moloi, M. Q. (2005, September). Mathematics achievement in South Africa: A comparison of the official curriculum with pupil performance in the SACMEQ II Project. In SACMEQ International Invitational Conference, International Institute for Educational Planning (pp. 28-30)

Moloi, M. Q., & Chetty, M. (2010). The SACMeQ III Project in South Africa. SACMEQ.

Mohohlwane, N.L. (2016). *The Contribution of Randomised Control Trials (RCTs) To Improving Education Evaluations for Policy: Evidence from Developing Countries and South African Case Studies*. A research report submitted to the Wits School of Education, University of Witwatersrand, in partial fulfilment of the requirements for the degree Master of Education

Motshekga, A. (2016). Statement delivered on 29 Jan 2016 following the meeting of the Council of Education Ministers' meeting on 28 January 2016.

Muller, J. (2004). Assessment, qualifications and the national qualifications framework in South African schooling. *Changing class: Education and social change in post-Apartheid South Africa*, 221-246.

Mullis, I. V., Martin, M. O., Minnich, C. A., Tanco, G. M., Arora, A., Centurino, V. A., & Castle, C. E. (2012). TIMSS 2011 encyclopedia. *Education Policy and Curriculum in Mathematics and Sciences. Boston: TIMSS & PIRLS Intenational Center, Lynch School of Education, Boston College*.

Mullis, I. V., Martin, M. O., Foy, P., & Arora, A. (2012). *TIMSS 2011 international results in mathematics*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Mweli, H., (2016). Remarks by Director-General on the Draft National Integrated Assessment Framework by the Department of Basic Education as reported to the Portfolio Committee on Basic Education on 8 November 2016 for the 2nd Quarter 2015/16 performance progress report. Downloaded 8 February 2017 from [https://pmg.org.za/committee-meeting/23599/]

National Center for Education Statistics (2015). Nation's Report Card resources. Downloaded 15 August 2016 from [http://nces.ed.gov/nationsreportcard/]

National Planning Commission. (2012). National Development Plan 2030: Our future–make it work. *Presidency of South Africa, Pretoria*, *1*.

Ndhlovu, T., Sishi, N., & Deliwe, C. N. (2006). A review of ten years of assessment and examinations. *Marking Matric, HRSC Press, Cape Town*.

Nuga Deliwe, C. (2004). An Assessment of ten years of education and training in South Africa. Report written for the Department of Education, Pretoria (2004).

Onsomu, E., Nzomo, J., & Obiero, C. (2005). The SACMEQ II project in Kenya: a study of the conditions of schooling and the quality of education. *Harare, Zimbabwe: SACMEQ*

Oreopoulos, P., & Salvanes, K. G. (2009). *How large are returns to schooling? Hint: Money isn't everything* (No. w15339). National Bureau of Economic Research.

Organisation for Economic Co-operation and Development. (2008). *Reviews of national policies for education: South Africa*. OECD Publishing.

Organisation for Economic Co-operation and Development. (2010). Strong Performers and Successful Reformers in Education: Lessons from PISA for the United States. Retrieved in August 2017 from [http://www.oecd.org/pisa/pisaproducts/46581323.pdf]

Organisation for Economic Co-operation and Development. (2013). *Synergies for better learning: An international perspective on evaluation and assessment: An*

international perspective on evaluation and assessment. South Africa. OECD Publishing.

Overton, T. (2011). *Assessing learners with special needs: An applied approach*. Pearson Higher Education.

Phelps, R. P. (2006). Characteristics of an effective student testing system. *Educational Horizons*, *85*(1), 19-29.

Phelps, R. P. (2012). The effect of testing on student achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21-43.

Poliah, R. R. (2001). Enhancing the Quality of Assessment in South African Schools. Conference Proceedings of the 27th IAEA Conference in Rio de Janeiro.

Poliah, R. R. (2011). *The management of the quality assurance of school based assessment at a national level in South Africa* (Doctoral dissertation).

Poliah, R.R. (2014). Twenty years of democracy in South Africa: a critique of the examinations and assessment journey. Unpublished document.

Ramatlapana, K., & Makonye, J. P. (2012). From too much freedom to too much restriction: The case of teacher autonomy from National Curriculum Statement (NCS) to Curriculum and Assessment Statement (CAPS). *Africa Education Review*, *9*(sup1), S7-S25.

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

Reddy, V. (2006). *Mathematics and science achievement at South African schools in TIMSS 2003*. HSRC Press.

Reddy, V., Prinsloo, C., Arends, F., Visser, M., Winnaar, L., Feza, N., ... & Ngema, M. (2012). Highlights from TIMSS 2011: the South African perspective. Human Sciences Research Council, Pretoria.

Reddy, V., Visser, M., Winnaar, L., Arends, F., Juan, A and Prinsloo, C.H. (2016). TIMSS 2015: Highlights of Mathematics and Science Achievement of Grade 9 South African Learners. Human Sciences Research Council, Pretoria.

Reeves, C. A. (2005). The effect of 'opportunity-to-learn' and classroom pedagogy on *mathematics achievement in schools serving low socio-economic status communities in the Cape Peninsula* (Doctoral dissertation, University of Cape Town).

Ross, K., Saito, M., Dolata, S., Ikeda, M., Zuze, L., Murimba, S., & Griffin, P. (2005). "The conduct of the SACMEQ III project." In *Onsomu, E., Nzomo, J., Obiero, C.,(2005). The SACMEQ II Project in Kenya: A Study of the Conditions of Schooling and the Quality of Education. Harare: SACMEQ. Available at http://www. sacmeq. org/sacmeqprojects/sacmeq-ii/reports*.

Rosenkvist, M. A. (2010). Using Student Test Results for Accountability and Improvement: A Literature Review. OECD Education Working Papers, No. 54. *OECD Publishing (NJ1)*.

Rouse, C. E., Hannaway, J., Goldhaber, D., & Figlio, D. (2007). *Feeling the Florida heat? How low-performing schools respond to voucher and accountability pressure* (No. w13681). National Bureau of Economic Research.

SABER. (2014). Systems Approach for Better Education Results Student Assessment rubrics. Retrieved 23 July 2016 from www.worldbank.org from: http://wbgfiles.worldbank.org/documents/hdn/ed/saber/supporting_doc/Background/SAS/Rubrics_SA.pdf.

SAB&T Deloitte. 2013. Technical report on Verification ANA (V-ANA) results 2013 for the Department of Basic Education. Dated 11 February 2014. Unpublished.

South African Press Agency. (2011). *Education failing: NAPTOSA. 28 June*. Retrieved 15 August 2016 from: http://www.news24.com/SouthAfrica/Politics/Education-failing-teachers-say-20110628

SADTU. (2011). *SADTU'S discussion document in response to ANA report*. Retrieved 15 August 2016 from: http://www.sadtu.org.za/docs/disc/2011/ana_report.pdf

SADTU (2014). SADTU: Over-emphasis on tests, assessments. Retrieved June 2016 from: http://www.news24.com/SouthAfrica/News/Sadtu-Over-emphasis-on-tests-assessments-20140928

Schiefelbein, E., & McGinn, N. F. (2008). *Learning to Educate: proposals for the reconstruction of Education in Latin America*. UNESCO, International Bureau of Education.

Schleicher, A. (2009). Securing quality and equity in education: Lessons from PISA. *Prospects*, *39*(3), 251-263.

Sen, A. (2003). Development as capability expansion. *Readings in human development*, 3-16.

Shepherd, D. L. (2011). Constraints to school effectiveness: What prevents poor schools from delivering results. *Programme to support pro poor policy development programme. Department of Economics, Stellenbosch University. PSPPD Project–April*.

Singh, A. (2015). *How standard is a standard deviation? A cautionary note on using SDs to compare across impact evaluations in education. Web log posting*. Retrieved 22 July 2016 from: http://blogs.worldbank.org/impactevaluations/how-standard-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations

Slavin, R. E. (2005). Evidence-based reform: Advancing the education of students at risk. Report Prepared for: Renewing Our Schools, Securing Our Future.

Snilstveit, B., Stevenson, J., Phillips, D., Vojtkova, M., Gallagher, N., Schmidt, J., Jobse, H., Geelen, M., Pastorello, M. & Eyers, J. (2015). *Interventions for improving learning outcomes and access to education in low- and middle-income countries: a systematic review*. International Initiative for Impact Evaluation (3ie).

South African Human Rights Commission (SAHRC) and the United Nations Children's Fund (UNICEF). (2014). South Africa Poverty traps and social exclusion among children in South Africa. Report developed by the Research on Socio-Economic Policy (ReSEP), Department of Economics, University of Stellenbosch.

Spaull, N. (2011a). A preliminary analysis of SACMEQ III South Africa. *Stellenbosch: Stellenbosch University*.

Spaull, N. (2011b). Primary school performance in Botswana, Mozambique, Namibia and South Africa. *SACMEQ III*.

Spaull, N. (2013a). Poverty & Privilege: Primary School Inequality in South Africa. International Journal of Educational Development. 33(2013) pp. 436-447.

Spaull, N. (2013b). South Africa's education crisis: The quality of education in South Africa 1994-2011. *Report Commissioned by CDE*, 1-65.

Spaull, N., & Taylor, S. (2012). *Effective enrolment–Creating a composite measure of educational access and educational quality to accurately describe education system performance in sub-Saharan Africa*. Stellenbosch Economic Working Papers 21/12.

Strauss, J.P. (1999). *Results of the Monitoring Learning Achievement Project. Discussion Document. Research Institute for Education Planning.* University of the Free State. Bloemfontein. Web site:http://www.education.gov.za/LinkClick.aspx?fileticket=mDkxzCGj2n8%3D&tabid=454&mid=404

Taylor, N. (2011). The National School Effectiveness Study (NSES): Summary for the synthesis report. *JET Education Services, Johannesburg*.

Taylor, N., Muller, J., & Vinjevold, P. (2003). *Getting schools working: Research and systemic school reform in South Africa*. Pearson South Africa.

Taylor, N., Van der Berg, S. & Mabogoane, T. (2013). What makes schools effective? Report of the National Schools Effectiveness Study. Cape Town: Pearson.

Taylor, N., & Vinjevold, P. (Eds.). (1999). *Getting learning right: report of the President's Education Initiative Research Project*. Joint Education Trust.

Taylor, S. (2011). *Uncovering indicators of effective school management in South Africa using the National School Effectiveness Study* (Doctoral dissertation, University of Stellenbosch).

Taylor, S., & Yu, D. (2009). The importance of socio-economic status in determining educational achievement in South Africa. Unpublished working paper (Economics) Stellenbosch: Stellenbosch University.

Umalusi (2004). Investigation into the Standard of the Senior Certificate Examination: A Report on Research conducted by Umalusi. Umalusi. Pretoria.

UNESCO. (2004). *EFA Global Monitoring Report 2005. Education for All: The Quality Imperative.* Paris: UNESCO Publishing. Chapter 1 retrieved Jan 15, 2015 http://www.unesco.org/education/gmr_download/chapter1.pdf

UNESCO. (2013). *Training Tools for Curriculum Development: A Resource Pack.* Geneva: UNESCO International Bureau of Education (UNESCO-IBE).

UNESCO. (2015a). *EFA* Global Monitoring Report 2015. *Education for All 2000-2015: Achievements and Challenges - Global Monitoring Report.* Paris: UNESCO Publishing.

UNESCO. (2015b). *Education 2030 Incheon Declaration and Framework for Action: Towards inclusive and equitable quality education and lifelong learning for all*. Paris, UNESCO.

Van der Berg, S., & Louw, M. (2007). Lessons learnt from SACMEQII: South African student performance in regional context. University of Stellenbosch, Department of Economics and Bureau for Economic Research Working Paper, 16(07).

Van der Berg, S. (2008). How effective are poor schools? Poverty and educational outcomes in South Africa. *Studies in Educational Evaluation*,*34*(3), 145-154.

Van der Berg, S. (2009). Fiscal incidence of social spending in South Africa, 2006. *Stellenbosch Economic Working Papers*, *10*.

Van der Berg S., Kruger J., Gustafsson M., Trawle G, Burger R, Folscher A, van Wyk C, Taylor S. (2012). Public Expenditure Analysis Report for South Africa, developed by Oxford Policy Management and the Research on Socio-Economic Policy (ReSEP), Department of Economics, University of Stellenbosch, funded by the United Nations Children's Fund (UNICEF) South Africa

Van der Berg, S., & Shepherd, D. (2010). *Signalling performance: Continuous assessment and matriculation examination marks in South African schools* (No. 28/2010).

Van Staden (2014). *Exploring possible effects of Differential Item Functioning on reading achievement across language subgroups: A South African perspective*. Paper presented at the World Education Research Association.

Vivalt, E. (2015). How much can we generalize from impact evaluations. *Unpublished manuscript*.

Wagner, D. A. (2010). Quality of education, comparability, and assessment choice in developing countries. *Compare*, *40*(6), 741-760.

Wasanga, P. M., & Ndege, J. G. (2012) Designing the assessment of learning outcomes to make positive impact on individuals and institutions. Experience of Kenya. Presentation by Kenya to the 38th IAEA conference in Kazakhstan.

Williams, J. H., & Engel, L. C. (2013). Testing to rank, testing to learn, testing to improve: An introduction and overview. *Research in Comparative and International Education*, *8*(3), 214-235.

Young, (2011). Daily Telegraph commentary. "This is conclusive proof that Labour claim to have improved schools while in office is utter nonsense. British schoolchildren plummeted in international league tables."

Zuma, J.G. (2011). *Republic of South Africa*: *State of the Nation Address 2011*. Retrieved 23 July 2016 from http://www.gov.za/state-nation-address-his-excellency-jg-zuma-president-republic-south-africa

Appendix.

Letter of authorisation from Department of Basic Education for this research study

basic education
Department:
Basic Education
REPUBLIC OF SOUTH AFRICA

Private Bag X895, Pretoria, 0001, Sol Plaatje House, 222 Struben Street, Pretoria, 0002, South Africa
Tel.: (012) 357 3000, Fax: (012) 323 0601, www.education.gov.za

Ref no: ODG-2439-29/01/2016
Enquiries: Ms M Matabane
Tel: 012 357 3658
Email: Matabane.a@dbe.gov.za

Ms C Nuga-Deliwe
Chief Directorate: Strategic Planning, Research & Coordination
Department of Basic Education
PRETORIA
0001

By email: Nuga.C@dbe.gov.za

Dear Ms Nuga-Deliwe

REQUEST FOR APPROVAL TO ACCESS DATA AND DOCUMENTATION FROM THE DEPARTMENT OF BASIC EDUCATION, FOR RESEACH STUDY ON MEASURING SYSTEM PERFORMANCE IN THE BASIC EDUCATION SECTOR

On 14 January 2016, the Department of Basic Education (DBE) received your research request to conduct data and document collection through accessing data, documents, reports, analysis and instrumental documentation indicating historical development, in measurement of learning and system performance quality as part of your Master in Education: Measuring System Performance in the Basic Education Sector, at the University of Witwatersrand.

The research request is approved on condition that you, as the applicant of the research data,

1. adhere to the conditions set in the research protocol document titled, "Guidelines for researchers in conducting research in the Department" (available on the DBE intranet and internet);
2. ensure anonymity of the data and maintain confidentiality of sensitive information contained within the ANA (Annual National Assessments), NSC (National Senior Certificate), SACMEQ (Southern and Eastern Africa Consortium for Monitoring Educational Quality), TIMSS (Trends in International Mathematics and Science Study), PIRLS (Progress in International Reading Literacy Study), and, Systemic Evaluations data sets; as well as the associated documentation; and
3. present the proposed acceptance letter from University of the Witwatersrand, to the RCME unit upon receipt.

Please be aware, you have been granted approval to access and utilise only the ANA, NCS, SACMEQ, and Systemic Evaluations, PIRLS and TIMSS data sets and associated documentation. Should you require additional data or information, you will be requested to resubmit a research request to the DBE.

As guided by the Department's Research Protocol, we deduce that you are aware of the ethical and legal responsibilities towards the research data and that you will protect the welfare and maintain the privacy and confidentiality of all data records provided; at all times;.

Please also note, we request that you share the findings of the research with the DBE at the conclusion of your research. Kindly supply the RCME unit with two copies for the attention of Director-General.

Yours sincerely

MR HM MWELI
DIRECTOR-GENERAL
DATE: 29/01/2016