Identifying Social Indicators for the BRICS using Public data

An investigation of the School Dropout Phenomenon in Brazil, India and South Africa

by

Krish Chetty

Thesis presented in fulfilment of the requirements for the degree of Master of Philosophy (Information and Knowledge Management) in the Faculty of Arts and Social Sciences at Stellenbosch University

Supervisor: Ms Heidi van Niekerk

March 2017

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

March 2017

Abstract

Identifying Social Indicators for the BRICS using public data:

An investigation of the School Dropout Phenomenon in Brazil, India and South Africa

Krish Chetty

Department of Information Science

University of Stellenbosch

Dissertation: Master of Philosophy (Information and Knowledge Management)

March 2017

The Brazilian, Russian, Indian, Chinese and South African (BRICS) Heads of State in 2014, in Fortaleza Brazil called for the closer cooperation of their statistical agencies and experts to promote the identification of common data methodologies that can be employed to analyse social indicators which measure a common set of challenges in their countries. The study examines the possibility of using data produced locally within Brazil, India and South Africa specifically to assess the singular but complex phenomenon of learners dropping out of school. Although the countries share a common challenge, the reasons behind the challenge differ based on the countries' varied backgrounds. In addition, each of the countries measure school dropout rates differently but in essence only considers the number of learners who dropout, whilst not describing the determinants of this dropout. This study employs Amartya Sen's Capability Approach to identify these determinants by identifying the central freedom affecting the learner, viz., the learner's real freedom to complete school and attain employment and an improved quality of life. This freedom is tested in terms of a Capability Set of functionings that learners aspire to attain or conduct, viz., being physically well, being financially secure, being mentally well, being taught in infrastructure of a suitable standard, being in a conducive home learning environment, travelling to school in a safe manner, feeling free to express themselves in school and lastly, effectively participating in school activities in a meaningful way. These broad functionings are further defined in terms of themes and sub-themes and thereafter datasets from the above mentioned 3 countries are identified in terms of questions that are appropriate to assess the performance of the country. However, the key additional step of this study is to qualify the selection of data variables per sub-theme in terms of the associated level of data quality. By applying data quality theory, a set of dimensions are identified, which are applicable to a data user working with a publicly released dataset. The selected datasets are checked in terms of relevance internationally and amongst Brazil, India and South Africa in terms of their data collection policy priorities. South Africa’s Statistical Assessment Framework was found highly useful, as the framework shared many of the identified data quality dimensions and assisted in developing the framework practically. In applying the newly constructed Public Data Quality Assessment Framework, the identified datasets were assessed in terms of the data quality dimensions and their level of data quality was rated. South Africa’s surveys produced by Statistics South Africa were rated strongest. Ultimately, relevant data can be sourced from the BRICS, however the variables identified are nuanced and pertain to the priorities of the countries. Greater effort is need to promote collaboration amongst the BRICS to produce comparable data, informed by common methodologies and data quality standards.

Opsomming

Identifiseering van Sosiale Aanwysers vir die BRICS deur die gebruik van data:

‘n Ondersoek na die skoolverlating fenomeen in Brasilië, Indië en Suid-Afrika

Krish Chetty

Departement Inligtingkunde

Universiteit van Stellenbosch

Proefskrif: Magister in die Wysbegeerte (Inligting- en Kennisbestuur)

Maart 2017

In 2014 het die Staatshoofde van Brasilië, Rusland, Indië en Suid-Afrika(BRICS), in Fortaleza, Brasilië, versoek dat hul statistiek-agentskappe en –kundiges nouer moet saamwerk ter bevordering van die identifisering van gemene data metodologië wat prakties aangewend kan word om sosiale aanwysers, wat 'n stel gemeenskaplike uitdagings meet, te analiseer. Die studie ondersoek die moontlikheid om data wat plaaslik in Brasilië, Indië en Suid-Afrika verwerf is, te gebruik om die spesifieke verskynsel van leerders wat skool te vroeg verlaat (vroeg uitval), te analiseer. Alhoewel bogenoemde 'n gemeenskaplike uitdaging is verskil die redes vir die verskynsel op grond van lande se uiteenlopende agtergronde. Daarbenewens meet elke land die skooluitvalsyfer anders en in wese word slegs die aantal leerders wat uitval in ag geneem, sonder om die redes daarvoor te beskryf. Hierdie studie gebruik Amartya Sen se “Capability Approach”, (Vermoënsbenadering) om die redes te identifiseer. Sen se benadering fokus op die identifisering van die sentrale vryheid van die leerder, nl. die ware vryheid van die leerder om sy/haar skoolloopbaan te voltooi en om n indiensnemingsvlak te bereik om sodoende 'n verbeterde lewenskwaliteit te bekom. Hierdie vryheid word gemeet in terme van n’ vermoë-stel van funksionaliteit wat leerders wil bereik, nl. om fisiek gesond te wees, om finansieël veilig te wees, om geestelik gesond te wees, om onderrig te word in 'n infrastruktuur van 'n geskikte standaard, om 'n bevorderlike leeromgewing tuis te hê, om veilig by die skool te kom, om vry te voel om hom/haarself by die skool uit te druk en laastens om effektief en sinvol deel te neem aan skoolaktiwiteite. Hierdie funksionaliteit word verder tematies en subtematies omskryf. Daarna word stelle data van bogenoemde lande geïdentifiseer deur middel van vrae wat geskik is om die lande se prestasie te evalueer. Die addisionele sleutelstap van hierdie studie is egter om die seleksie van dataveranderlikes per subtema te kwalifiseer met betrekking tot die verbandsvlak van kwaliteit data. Deur n datakwaliteitsteorie toe te pas, word 'n stel dimensies geïdentifiseer, wat bruikbaar is vir 'n data gebruiker wat te doen het met publieke publikasies van data. Die geselekteerde stelle data word geverifieer in terme van die relevansie daarvan internasionaal en is ook geweeg teen Brasilië, Indië en Suid-Afrika se dataversamelings-beleidprioriteite Suid-Afrika se Statistieke Assesseringsraamwerk is baie nuttig, aangesien die raamwerk baie van die geïdentifiseerde data se kwaliteitsdimensies kon deel en hulp verleen met betrekking tot die praktiese ontwikkeling van die gebruikte raamwerk. Die nuut-saamgestelde Openbare Datakwaliteitsraamwerk is gebruik om die geïdentifiseerde stelle data te evalueer in terme van die datakwaliteitsdimensies en die gradering van die kwaliteitsvlak van die data. Suid-Afrika se opnames, gedoen deur Statistieke Suid-Afrika, is die hoogste gegradeer. Uiteindelik kan relevante inligting van BRICS verkry word. Die veranderlikes is egter genuanseerd en het betrekking op die land se prioriteite. Meer pogings is nodig om samewerking onder die BRICS lande aan te moedig, ter bevordering van die inwin van data deur gemene metodologië en datakwaliteitsstandaarde.

Acknowledgements

I would like to firstly thank and express my sincere appreciation to Ms Samantha Petersen for the many months of constant support during the period of this project.

Secondly, I would like to thank my mother Dr Carmel Chetty and my manager at the Human Science Research Council BRICS Research Centre (HSRC BRC) Dr Jaya Josie, for their assistance with editing and guidance in structuring this document.

Thirdly, I would like to recognise the efforts of Dr Rumi Aijaz from the Observer Research Foundation (India) and Mr Jefferson Pecori Viana from the NEBRICS Centre at the Federal University of Rio Grande do Sul in assisting to find relevant datasets from India and Brazil respectively.

Lastly, I would like to whole-heartedly thank my supervisor Ms Heidi Van Niekerk for her always timely and thoughtful feedback and guidance during the length of this project.

Contents

Part I: Theoretical Framework

Part II: Empirical Framework

Part III: Findings

List of Tables

List of Figures

List of Appendix Tables

List of Appendix Figures

Chapter 1

Introduction

This investigation into the School Dropout Phenomenon affecting Brazil, India and South Africa is cased within the need to identify suitable data sources, produced within Brazil, Russia, India, China and South Africa (BRICS) for the reporting of relevant social indicators. This study provides a framework for selecting publicly released datasets to assess a singular phenomenon affecting the 3 selected countries of the study, viz., learners who drop out of school. This is intended to illustrate to the BRICS nations how such a study can be expanded to cover the broader needs of the social sector across all 5 countries and how to assess the degree of data quality inherent to each relevant data collection instrument. In brief, this assessment of data quality investigates the possibilities of reporting on BRICS trends using data produced internally within BRICS countries.

1.1. Background

The Fortaleza Declaration produced by the BRICS' Heads of State in 2014 following the Heads of State Summit in Fortaleza, Brazil made a clear statement calling for the BRICS countries to establish a joint methodology for the purpose of reporting on social indicators. Action point seven in the declaration states:

> To better reflect the advancement of the social policies of the BRICS and the positive impacts of its economic growth, we instruct our National Institutes of Statistics and the Ministries of Health and Education to work on the development of joint methodologies for social indicators to be incorporated in the BRICS Joint Statistical Publication. We also encourage the BRICS Think Tanks Council to provide technical support in this task. We further request the BRICS National Institutes of Statistics to discuss the viability and feasibility of a platform for the development of such methodologies and to report thereon (BRICS 2014).

The data that is reported on or referred to in research outputs produced by BRICS researchers in the existing official BRICS fora is often based on the existing data repositories and publications made available by the likes of the World Bank, International Monetary Fund (IMF), Organisation for Economic Co-operation and Development (OECD); and the various

United Nations (UN) bodies. Consequently, the research outputs that are produced, which informs BRICS strategy and policy is based on data and information that is often contentiously viewed by the BRICS nations and possibly differs from the figures reported on locally within the BRICS nations. It is these concerns that have led to the call by the BRICS Heads of State for their statistical agencies to develop joint methodologies for social indicators.

The BRICS member states are making a concerted effort to reduce their dependence on organisations such as the World Bank. The formation of the BRICS New Development Bank that has come into being with an initial investment of $100 shared amongst the nations (BRICS 2014) underscores the emphasis that has been placed on independence and development in BRICS countries. The purpose of the bank is to fund long term investment in sustainable development and general infrastructure requirements that have not been supported by the existent development bank facilities (Griffith-Jones 2014). In light of the development challenges faced by the BRICS, identifying the priority areas are a major concern. Such priorities are only identifiable using data that is available. This data informs BRICS policy. Therefore, the need for common agreeable data will help drive common agreeable BRICS actions.

Researchers' current reliance on the existent set of information made available by international bodies such as the World Bank is due to a perceived notion that such data and information are reliable and of a high quality. Ben Kiregyera (2015, pp 39-41) suggested that the reliance on international organisation's databases was often due to neglect or the lack of development of databases within African countries in the 1970s and 1980s whilst in this period, organisations such as the World Bank strengthened their own databases on these countries (Kiregyera 2015, pp 39-41). However, as mentioned, the accuracy of such data is often disputed in BRICS circles.

The controversy surrounding the use of World Bank data relates to the manner in which their data models are compiled and their decisions taken to impute missing data values or in the definitions applied to aggregate data obtained from multiple sources with incompatible methodologies. The World Bank itself cautions users when using data across multiple sources with the following quote provided on their website, "*Differences in timing and reporting practices may cause inconsistencies among data from different sources, so users should be cautious when combining these data*" (The World Bank n.d.). Kiregyera (2015, p 51). suggests that data from international bodies is often initially based on nationally collected data but is then processed using models to allow the data to be comparable across countries. The manner in which these manipulations are carried out often leads to disputes amongst countries. An anecdotal example of this controversy of using data from international organisations such as the World Bank is illustrated in the World Bank's publication of an unemployment figure for Zimbabwe at 4%. Zimbabwe's unemployment rate was a contentious and politically sensitive topic. Zimbabwe's political parties both presented conflicting figures for unemployment (greater than 60%), whilst the World Bank's figure was reported on a modelled estimate that was based on data sourced over a decade ago and the World Bank's imputation techniques had not factored in the economic crash of Zimbabwe (Chiumia 2014).

For the purpose of this study the ambit of social indicators as requested to study by the BRICS Heads of State is too broad and requires a narrower focus on a singular concern affecting each of the nations. One such area that requires examination is the phenomenon of learners dropping out of the school system and the broader contextual concerns related to the school, household and community which leads to such a choice. School dropouts are prevalent in South Africa, Brazil and India and are not clearly defined nor are the determinants for such occurrences reported on by the government departments responsible for education considering that a comprehensive indicator of dropping out is far more complex than noting the reduction in learner enrolments over time. Cardoso and Verner (2006) identified that there are various 'push-out' factors which contribute to school dropouts in Brazil such as early parenthood, child-labour and poverty (Cardoso & Verner 2006). In South Africa, great attention is placed on increasing pass rates whilst there is scarce reporting on the phenomenon of school dropouts (Rademeyer 2014). Calculations performed by the South African Institute for Race Relations in 2015 found that in South Africa approximately only 50% of learners that enrolled in grade 1 reached their final year and 30% of those learners pass the final year examination (South African Institute of Race Relations 2015). In India, a report by the National University of Educational Planning and Administration (2011) identified the dropout problem as pervasive to the Indian education system and identified factors such as poverty and family and domestic problems which create an environment in which education is devalued (Chugh 2011). Reddy and Sinha, in 2010, raised a similar perspective to Cardoso and Verner in their analysis of dropouts in India and also referred to the phenomenon as a pushout instead of a dropout from the perspective of the learner. By referring to issues such as poverty, child labour, household decisions, school quality and the curriculum, amongst others, the authors noted that the majority of the factors relate to macro-economic pressures which tend to force the learner out of school instead of simply assuming the learner drops out due to an unwillingness to persevere in school (Reddy & Sinha 2010). Although each country is affected by a common problem, their diverse backgrounds influences the learner's rationale for dropping out of school.

The diverse nature of the BRICS also impacts their data collection and reporting approaches. The BRICS national statistical agents have been mandated to develop a common reporting methodology but are challenged by their differences. Selecting the optimal set of comparable indicators for the BRICS can be difficult especially in terms of identifying commonalities across the nations. In Arruda, Slingsby, Ustyuzhantseva and Nafey's (2015) discussion of the state of the education systems at the 7th BRICS Academic Forum they found that:

> Despite a vast amount of data existing at the national level, there is still a challenging lack of comparable data series on the BRICS (as exemplified by the lack of data on enrolment for Brazil, repetition for South Africa and literacy for India). This is a major barrier to cooperative policymaking at the inter-regional level (Arruda et al. 2015).

The authors suggested that the BRICS make an effort to identify how to standardise their data reporting using a strategy more appropriate to their particular needs in a manner different in approach to the one conducted by an organisation such as the United Nations Educational, Scientific and Cultural Organization (UNESCO).

In order to address Arruda et al's (2015) findings, the BRICS must craft a strategy for data collection targeted to the needs of the BRICS countries. This study offers some guidance on the data collection efforts required to address a single sub-component of the needs for Social Indicators in the BRICS by focussing on identifying the key factors which describe determinants of dropping out of school, specifically in Brazil, India and South Africa. This study argues that in order to select the appropriate set of indicators for the BRICS, these nations will require a two-pronged approach. This requires a focus on

a) A conceptual approach to identifying key themes in relation to the school dropout phenomenon, informed by the Capability Approach of Amartya Sen and
b) A technical approach informed by issues of data quality as discussed by Shazia Sadiq and Martin Eppler amongst other notable Data Quality authors.

This study argues that for the BRICS context; the incorporation of the data quality dimension is equally valuable when selecting data from BRICS data sources.

1.2. Research Problem

Against the above-mentioned background, the hypothesis of the study has been formulated as follows:

Publicly available data produced by the BRICS can be used as a source for alternate social indicators within the BRICS. Despite the current tendency of sourcing data from international data providers, there is sufficient data of a high quality that is available to inform research within these countries.

The research approach followed within this study describes the phenomenon of school dropouts across Brazil, India and South Africa and can be expanded to include the remaining BRICS countries viz. the Russian Federation and China and the wider aspects of social service provision.

1.2.1. Research Questions

To test this hypothesis, the following four questions were identified:

The primary question of this study (Question 1) relates to the need of the BRICS and specifically Brazil, India and South Africa to produce indicators in order to reduce their dependence on foreign sources. Therefore, the key word raised within the hypothesis is 'alternate.' Furthermore, the relevance of the data must be assessed in terms of what learners value the most as well as in terms of the associated data quality of the dataset. One must therefore determine how to assess indicators' relevance in light of the core concerns of the phenomenon as experienced in the three selected countries as well as in terms of the level data quality that the dataset exhibits.

Question 1	Can public data produced within Brazil, India and South Africa be used to describe factors leading to school dropouts?
Question 2	What framework can be adopted to identify data that describes factors leading to school dropouts?
Question 3	Which attributes within datasets can be utilised in defining the capability approach relevant themes describing the school dropout phenomenon?
Question 4	What are the data concerns, if any, that impede the use of data that describes the factors driving the school dropout phenomenon?

Table 1: Research questions of the study

The manner in which the indicators are selected requires a defensible and theoretical framework (Question 2). As is identified in the literature review, local studies often refer to indicators without theoretical reasoning determining why such a selection of indicators was made. The frameworks that are adopted will be used to assess issues raised within Sen's Capability as well as incorporating Data Quality concerns discussed within Sadiq's Data Quality Handbook (Ge et al. 2013) and by Martin Eppler in 2006. The combination of approaches will help ensure that the selection framework identifies useful and quality data. The framework must identify the core concerns and areas of commonality across the three countries. In addition, where there are concerns unique to a particular country this should be identified and the framework must cater for such occurrences.

Publicly available datasets produced within the basic education sectors of Brazil, India and South Africa that describe school attendance and the supporting environmental concerns faced by learners will need to be identified and thereafter assessed using the frameworks discussed in Question 2. Where an indicator is selected, the matching datasets that informs the creation of the indicator needs to be identified. Care will need to be taken to ensure that the indicators produced are based on data with similar methodologies in terms of data collection. As the framework defined in Question 2 will be used to assess the multitude of data sources available, the application of the framework will help identify which datasets are useful or not in terms of the Capability Approach as well as help determine whether suitable data quality practices were implemented to control the quality of the output data (Question 3).

Through the application of the dataset selection framework, the various identified datasets from Brazil, India and South Africa will be evaluated. If the data identified is of a poor quality, such concerns must be exposed as part of the findings in order to support the future assistance in the refinement of the dataset's data collection strategy (Question 4).

1.3. Applying the Capability Approach to analyse the School Dropout Phenomenon

Socio economic factors affecting learners are central to understanding the cause of the school dropout phenomenon, however, the identification of suitable socio-economic factors can be complicated. Sen's Capability approach is useful in the manner it frames access to education. Melanie Walker (2005a) finds that the Capability Approach provides a means to establish what freedoms a person is realistically able to achieve as opposed to simply considering the resources a person has accumulated or has access to. Furthermore, she notes that being educated has a direct bearing on additional freedoms one aspires to attain. Considering that capability can be interpreted as one's real freedom to achieve one's aspirations, one must ask whether such valued goals are within their means to reach. In order to attain this lifestyle, the quality of one's education experience must be assessed in terms of simple and complex functionings such as being well-nourished or having attained an adequate state of numeracy or literacy to participate in the learning process.

The popularity of the Capability Approach also follows from the numerous uses of it in the Human Development field. The approach lends itself to practical implementation following Sen's (2000, p75) argument for the identification of real freedoms which a person can achieve due to their particular circumstances or deprivations that one individually faces. These real freedoms can be expressed in terms of an enumerated set of valued states (functionings) that the person seeks to acquire or perform. Each enumerated functioning can thereafter be represented in terms of a measurable indicator which describes the level of attainment of the particular functioning. Sen believed that such an approach of identifying and measuring a collection of one's attainment of their relevant aspirations is much more beneficial to welfare economics than the traditional commodity based analysis which does not recognise the living conditions of individuals. It is for this change in focus to the needs of the disenfranchised in the field of welfare economics that won Amartya Sen the Nobel Prize in Economic Sciences in 1998.

The Capability Approach has since been operationalised in various efforts, most notably in the development of the Human Development Index (HDI) by Sen and Mahbub ul Haq for the United Nations Development Programme (UNDP) (United Nations Development Programme 1990 p10). Their development of the composite index was produced in an attempt to identify the most valued measures of human well-being that could thereafter be expressed in a single composite value, by aggregating the various identified measures together. The idea behind the composite index is to reflect both the pluralist nature of the many dimensions of human development, but, also reflect this in a single measure that could be used to rank development concerns and expose the shortcomings of the commodity approach's preferred measures such as GDP. Although the HDI was a breakthrough initiative and was adopted by the UNDP, the approach also has its critics. Developing a composite index for all countries in the world results in various limitations, as trade-offs were made to ensure a wider coverage of datasets. Thus the choices of indicators selected as well as the level of data quality used, is often criticised. It was Sen's opinion though that the benefit of the approach was also to begin the conversation pertaining to the development challenges faced by each country and thereafter explore in greater detail the valued functionings and real-freedoms that marginalised communities faced.

Within the education space, the work of Elaine Unterhalter, Rosie Vaughan and Melanie Walker (2015) has been crucial in applying the broad Human Development perspective to the needs of the education sector specifically. Unterhalter et al. has noted that one of Sen's more valued traits of his approach is to promote pluralism when describing the development challenge. Unterhalter et al. therefore examines in greater detail what the most valued sets of functionings were required in the education sector. This included the identification of factors such as class participation, developing vocational skills, numeracy, etc. Despite the narrower focus to education, education researchers are hesitant as to state whether the identified list of functionings are comprehensive. The authors thus, highlights the need for an even narrower focus on education related matters.

In view of Unterhalter et al.'s challenges, the essential factors which adequately describe the learner dropout phenomenon requires a deep and exhaustive review of current school dropout literature in the three countries that the study is applicable to. In South Africa, although, access to education is identified as a basic human right, a question remains whether the socio-economic conditions a learner finds themselves in actually guarantees or enables this right. Spaull (2013) expresses the challenges experienced in South Africa's education system as:

> This substandard education does not develop their capabilities or expand their economic opportunities, but instead denies them dignified employment and undermines their own sense of self-worth. In short, poor school performance in South Africa reinforces social inequality and leads to a situation where children inherit the social station of their parents, irrespective of their motivation or ability (Spaull 2013 p 9).

1.4. Assessment of Data Quality

Amongst the various criticisms of the Human Development Index and other efforts to operationalise the Capability Approach, a key concern that emerged regarded the quality of the datasets that were used to describe the valued functionings and real freedoms experienced by the targeted audience of their respective studies. Whilst various authors have applied the Capability Approach in a wide range of fields be it political, economic or human development, none have supplemented their approach with a data quality assessment to test the datasets that were chosen. The approach followed in this study is to examine the core elements of what constitutes data quality as valued by users of public data released in Brazil, India and South Africa. Through an examination of data quality theory, data quality dimensions which affect the data user are identified. These dimensions form the basis for testing the data quality of selected data used to describe a particular sub-theme amongst the various functionings of the school dropout phenomenon Capability Set.

Data quality is often examined from the perspective of the managing organisation and not particularly from the point of view of the end user. Shazia Sadiq, the editor of the Handbook on Data Quality (Ge et al. 2013) presents data quality from 3 crucial perspectives, the Organisational, the Architectural and the Computational. The organisational perspective is based on the concerns of organisation management, the architectural perspective caters to the concerns of operational staff and the computational perspective describes data quality in terms of the needs of defining the system. These concerns are combined with dimensions identified by Martin Eppler (2006) who considered data quality also from a range of perspectives, viz., the Community, Product, Process and Infrastructure levels. Through the application of data quality factors from Sadiq, Eppler and others, allows one to identify relevant data quality dimensions. Ultimately, the dimensions that are selected are done so with the user's requirements in mind.

From the analysis conducted, it is noted that data quality is a subjective concept that is multidimensional in nature similar to the valued functionings relevant to learners that dropout from school. By reviewing the data quality dimensions through the lens of the data user, only dimensions that the end user could assess based on information released to the public is identified. Furthermore, the theory is assessed in terms of the national priorities discussed by the statistical institutes in terms of data quality. South Africa's Statistical Assessment Framework (SASQAF) is found to be quite useful when applying the framework to assess the quality of public data. Interestingly, the SASQAF shares many of the data quality dimensions identified within the literature such as accuracy, clarity, applicability, comprehensiveness, currency, etc. This alignment has assisted in the development of the Public Data Quality Assessment Framework (PDQAF) which is discussed in greater in detail in Chapter 7.

By applying the PDQAF, ordinary data users can perform a data quality assessment on the available data. The key finding that emerges is that in order for the data user to develop an informed assessment of data quality, the data produced must provide support documentation which expresses how data quality is promoted by the organisation. As the framework positions the user's requirements centrally, the data providers also obtain an assessment of how the public perceives the quality of their output from this study.

1.5. Research Design and Methods

The research design followed in this study pertains to the development and application of two frameworks that refer to the School Dropout Capability Assessment Framework (SDCAF) and the Public Data Quality Assessment Framework (PDQAF). The figure below describes the steps followed in structuring this study. In terms of outputs of the study, 4 deliverables are produced, viz., (1) the SDCAF, (2) the PDQAF, (3) the identification of relevant data sets for each valued functioning and (4) the data quality assessment of the identified datasets from Brazil, India and South Africa. These 4 outputs are thereafter used to produce the findings of this particular study.

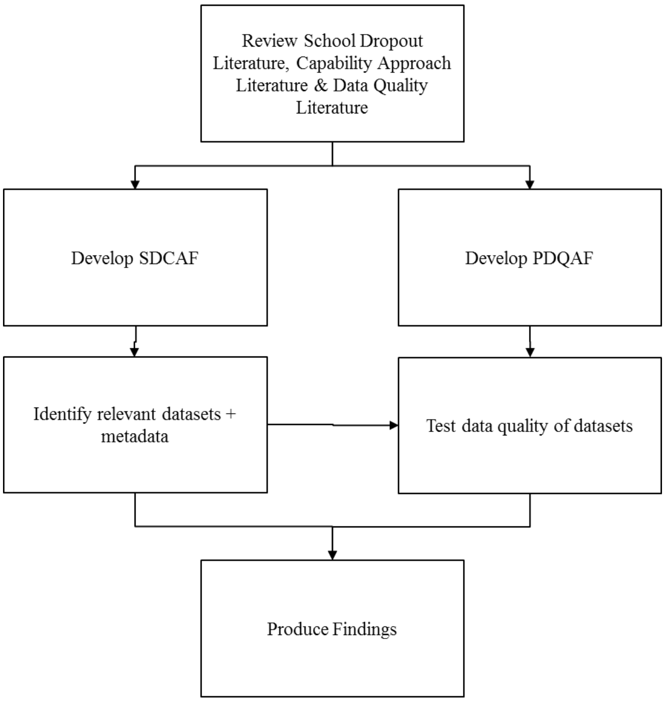

Review School Dropout Literature, Capability Approach Literature & Data Quality Literature
Develop SDCAF
Develop PDQAF
Identify relevant datasets + metadata
Test data quality of datasets
Produce Findings

Figure 1: Research Design Workflow

The study begins with an exploration of the current school dropout literature. Reasons for school dropouts are discussed as well as the current trends in the education sector within Brazil, India and South Africa. The notable factors which drive school dropouts are highlighted. This literature is thereafter analysed in terms of Amartya Sen's (1992, 1999, 2000, 2009) Capability Approach which provides an opportunity to contrast a learner's aspirations in life against their actual real freedoms due to environmental factors and limitations they face. The Capability Approach provides a structure which supports the enumeration of multiple factors which influence a learner's choice to drop out of school. These various factors are identified in existing implementations of the Capability Approach as well as in literature from the three countries which discuss the various education related challenges which the countries face. Furthermore, techniques presented by Alkire, Araina, Johnson, Naveed, Robeyns, Santos, Spence, Unterhalter and White (2009) and Comim, Qizilbash and Alkire (2008) are recognised as options for operationalising the Capability Theory. These techniques together with the literature are adapted to produce the SDCAF.

The structure of the SDCAF is based on a series of functionings which make up the school dropout Capability Set. Each functioning presents a broad aspiration that learner's value. Each functioning is further decomposed in terms of themes and sub-theme to provide a greater depth in granularity to assist in identifying specific relevant indicators from each country analysed in the study. Each sub-theme represents a broad space for numerous potential indicators depending on what data is produced within each country.

Together with the identification of useful datasets for each sub-theme, the identified dataset is assessed in terms of data quality. At this point, the PDQAF is utilised to assess the selected datasets. The PDQAF is informed by data quality studies by researchers such as Sadiq (2013), Eppler (2006), Loshin (2001), the International Monetary Fund (IMF) (2003, 2010, 2012) and Statistics South Africa (2008, 2010a, 2010b) are reviewed. Using the combination of resources, the PDQAF is produced and is thereafter used to test the data quality of each identified dataset. The findings of this assessment are presented in Appendix 2.

The data quality theories will help evaluate whether the identified datasets are of a suitable level of data quality for analysis and reporting based on the information that is made available to the public. The combination of the SDCAF and PDQAF allow the data user to broadly identify the extent to which data are available to describe each functioning's sub-theme per country and assess whether the selected datasets adequately describe the key sub-themes and whether each dataset is of a high data quality standard. With these 2 frameworks applied in concert, the resultant analysis of the three countries will inform the findings produced in this study.

1.6. Motivation for this study

The crux of this study is to present an example to the BRICS in terms of how to identify a core set of indicators which can be used to effectively monitor the common social challenges which affect the BRICS. The chosen approach as discussed in the research design section above is to identify a single challenge affecting a sub-set of the BRICS (Brazil, India and South Africa) and to identify what data is available to assess a collection of aspirations which describes this particular challenge. The example chosen in this study is to explore the range of aspirations of learners who choose to drop out of school.

Dropping out of school was chosen as access to education is a critical gateway to an improved quality of life as education equips people with a range of competencies which are necessary to become a productive member of society. The Millennium Development Goals (MGDs) for example, target access to schools as it is critical in developing a person's cognitive and non-cognitive skills (Pritchett 2004). However, the MDGs do not identify the determinants for leaving school. Whilst understanding that simply maintaining school attendance does not ensure the learner's development, the lack of access deprives the learner the opportunity of attaining such an enabling capability.

As Pritchett notes, the international development goals do not delve into sufficient depth which allows countries to better understand the determinants for learner's dropping out of school. In this light, the capability approach was identified to help elicit the reasons for a seemingly counter-intuitive behavioural choice. In order to quantify this practice and the reasons to do so, relevant data must be selected and its suitability for use must be assessed. The approach of the study is to assess the quality of data produced within Brazil, India and South Africa. It is the hypothesis of this study, that data in the BRICS can be used for reporting on the social challenges which beset these countries. There is a need to better understand how this data has been collected and to find the commonalities in existing data collection instruments utilised within these countries.

The assessment of available data describes the extent to which the learner's aspirations are met and the associated quality of this available data. This will identify which core concepts of this particular challenge are adequately described and whether the data available is suitable for research and general reporting. The findings of this study will illustrate the feasibility and limitations of using BRICS data to report on common BRICS challenges.

1.7. Limitations of the School Dropout Rate

Branson, Hofmeyr and Lam (2014) reviewed the progression of learners through the school system in South Africa and cautioned the South African Education Department of the high rate of grade repetition and the low numbers that completed the final year of school study. The South African Department of Basic Education identified in 2009 that the average grade repeater rate was 9% of all learners, whilst the average dropout rate per grade was 4% in 2007/08 (Department of Basic Education 2011). The presentation of these low year on year drop out percentages, tends to mask the scale of the school dropout problem in South Africa. Whilst 69% of South African learners reach the penultimate grade in the school system, only 39% of these learners successfully pass and matriculate from secondary school. Thus the majority of learners that leave the school system are semi-skilled or unskilled and therefore are not able to compete for the high productive jobs in the South African labour market.

In Brazil, the School Abandonment Rate, as it is referred to, is calculated as the percentage of learners who have stopped attending school as at the date that the school attendance census is carried out (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP n.d.). The calculation can be applied per grade, per school and across the various municipalities of Brazil. However, as the calculation is represented per grade, it neglects to reflect the accumulated problem that is experienced as the learner progresses within the school system for the duration of their school going experience. The calculation is as follows:

$$Abadonment\ Rate = \frac{\text{ABA}}{(\text{APR+REP+ABA})} \times 100$$

Where:

- ABA = Expected registrations not in attendance
- APR = Number of approved registrations
- REP = Returning approved learner

In India, as Raju and Singh (2011) explain, the dropout rate as measured by the Ministry of Human Resource Development quantifies the number of learners that dropout between beginning class of the stage under study and beginning class of the succeeding stage. The stages referred to in India are groupings of school grades or levels. There are five stages, the primary stage, the upper primary stage, the elementary stage, secondary stage and senior secondary stage. The dropout at upper primary level is calculated by subtracting the dropouts at the primary level from the elementary dropout numbers. Where the calculation is reflected as negative, the upper primary level dropout number is reported as zero. The primary stage includes grades I to V (the Indian school system uses roman numerals), the upper primary stage includes grade levels VI to VIII and the elementary stage includes grade levels I to VIII, the secondary stage refers to levels IX and X whilst the senior secondary stage refers to grade levels XI to XII (Government of India Ministry of Human Resource Development 2004). The calculation in India spans across grades and years requires the collection of data for the duration of the all learner's attendance in the school system for every year of attendance. The benefit of calculating in these terms allows one to review the cumulative effect of dropouts across a particular stage of schooling. The calculations for this rate at the various stages are:

$$Primary\ Dropout\ Rate\ (year\ t+4) = \frac{\text{Enrolment in Grade V in Year t+4}}{\text{Enrolment in Grade I in year t}} \times 100$$

$$Elementary\ Dropout\ Rate\ (year\ t+7) = \frac{\text{Enrolment in Grade VIII in Year t+7}}{\text{Enrolment in Grade I in year t}} \times 100$$

$$Upper\ Primary\ Dropout\ Rate\ (year\ t) = Elementary\ Dropout\ Rate\ (year\ t) - Primary\ Dropout\ Rate\ (year\ t)$$

$$Secondary\ Dropout\ Rate\ (year\ t+2) = \frac{\text{Enrolment in Grade X in Year t+2}}{\text{Enrolment in Grade IX in year t}} \times 100$$

$$Senior\ Secondary\ Dropout\ Rate\ (year\ t+2) = \frac{\text{Enrolment in Grade XII in Year t+2}}{\text{Enrolment in Grade XI in year t}} \times 100$$

In South Africa, the official reported learner dropout rate follows the UNESCO methodology which provides a year on year analysis regarding the change in enrolment per grade. The definition is the *"proportion of pupils from a cohort enrolled in a given grade at a given school year who are no longer enrolled in the following school year"* (Department of Basic Education 2011). Unlike India and similar to Brazil, South Africa uses the year on year calculation which again does not reflect cumulative effects of learner dropouts. This calculation is represented as:

$$Learner\ Dropout\ Rate\ (per\ Grade\ g+1)(Year\ t+1) = \frac{\text{Enr}-\text{Enr1}}{(\text{Enr})} \times 100$$

Where:

- Enr1 = Learners Enrolled in Grade g + 1 of Age Cohort x in year t +1
- Enr = Learners Enrolled in Grade g of Age Cohort x in year t

Dieltiens and Meny Gibert's (2009) analysis of school dropout has lead the authors to believe the causes of the phenomenon are broader than absolute poverty. Whilst poverty is a central factor, one must also consider the total cost of education which is inclusive of transport, school uniforms, stationery and the opportunity cost of the child's labour. In addition to these limitations in respect of household experience, Dieltien and Meny-Gilberts note that poverty does not explain all issues as in South Africa, despite high percentages of learners living in poverty in Grades R to nine, only a small percentage were not attending school. The authors find that one must consider a multidimensional model of poverty which includes income as well as other concerns such as access to basic services and the person's real ability to participate in society. Furthermore, they argue that a learner's attendance when in the non-compulsory school grades (ten, eleven and twelve) is impacted strongly by the quality of education they receive which is a much more complex set of issues. Firstly, the learner's accumulated state of learning from the previous grades may not be sufficient to participate in the higher school grades, whilst the learner may also find it difficult to contend with a poorer level of teaching. These difficulties combined with household difficulties and community factors indicate a wide array of factors which pressure learners to leave school. As a parent quoted within the author's essay stated, "… teachers in our schools don't teach our children and they don't care about our children's future. Our children don't feel free to go to school because of the bad things that happen in our community" (Dieltiens & Meny-Gilbert 2009 p48).

School dropout is often referred to as school pushout, which underlines the point that the factors that lead to a learner's leaving of school is not necessarily the choice of the learner but rather becomes unavoidable due to their circumstances stemming from their environment and experiences within the school. In this regard a simple presentation of reducing school attendance figures does not highlight the complexity of factors that belie the learner attendance reduction.

1.8. Definitions of Key Terms and Acronyms

BRICS	**Formation of 5 countries, working together in various areas of international relations and cooperation, viz., Brazil, Russian Federation, India, China and South Africa**
Capability	Capability refers to the actual real opportunity that an individual has to accomplish a task that the person truly values, given their personal circumstances.
Capability Set	The Capability Set is described as a vector or collection of 'functionings' which are identified as valuable 'beings' and 'doings' that a person could achieve.
Data Collection Instrument	A data collection initiative carried out by particular organisation to capture responses using a survey or is based on internal registers used in the operations of the organisation itself.

Dataset	A dataset is a subset of available data which is extracted from a data collection initiative. The degree of the granularity differs widely and is used to refer to a short table or the results of an entire survey.
Functioning	A functioning represents the choices or actions that individuals value, although may not be specifically attainable. Functionings are often referred to as a state of being or doing performed by an individual.
PDQAF	The PDQAF refers to the Public Data Quality Assessment Framework which is developed within this study to assess publicly released data. For further details, see Chapter 7.
SDCAF	The SDCAF refers to the School Dropout Capability Assessment Framework produced within this study to identify the key functionings, themes, sub-themes and thereafter related public data that can be used to describe the various components of the School Dropout Phenomenon

Table 2: Definitions of key terms and acronyms used frequently within the study

1.9. Thesis Outline

This thesis consists of 3 sections. The first part presents the theoretical basis of the study and groups together the various components of literature that are analysed as part of the Theoretical Framework. The second part of the thesis outlines the Empirical Framework which details the methodology and the construction of both School Dropout Capability Framework and the Public Data Quality Assessment Framework. The third past of the thesis includes the findings from the application of the framework and the conclusion of the study. Lastly the appendices present additional supporting information relevant to the application of the frameworks.

Part I of the study groups together the various components of literature that are analysed as part of the Theoretical Framework of the study. Chapter 1 unpacks the concerns related to learners that decide to dropout out of school within the 3 countries of this study. Chapter 3 breaks down the theory of the Capability Approach of Amartya Sen and discusses how the theory can be translated into an applied framework citing the various applications of the theory by notable scholars, paying specific attention to the application within the education sector and thereafter identifying steps that can be used to produce school dropout specific framework. The formation of this framework is crucial in working towards answering question 1 of the study which concentrates on identifying which datasets can be used to describe school dropouts in

Brazil, India and South Africa. However, to identify the dataset, first the most relevant set of themes that are linked to the key functionings, are first identified. Chapter 4 thereafter presents the key theoretical perspectives of data quality and by interpreting the theory through the lens of a data user, the most pertinent set of data quality dimensions are selected which are used to produce the foundations of the PDQAF.

Part II of the study outlines the Empirical Framework adopted within the study. Firstly, Chapter 5 describes the methodology that was followed in conducting the study. This includes the explanation of how the various assessments of literature are utilised in defining the SDCAF and the PDQAF. The methodology further describes the steps that are followed in developing the SDCAF and PDQAF. Chapter 6 and 7 thereafter outline the 2 frameworks that are created to identify the sets of data relevant for evaluating school dropouts in Brazil, India and South Africa. The application and assessment produced using the SDCAF thereafter directly answer the primary question of this study (Question 1) pertaining to relevant datasets to describe school dropouts. The 2 frameworks that are discussed in Chapter 6 and 7 directly answer Question 2 of the study pertaining to what frameworks can be adopted to identify data which describes determinants of school dropouts. Where the SDCAF is used to identify all potential sets of data, the PDQAF assists in shortlisting these datasets based on the assessment of data quality. Datasets of higher data quality are recommended over those of lower data quality and in the absence of data alternatives, cautions are provided to inform data users of the data quality concerns.

Part III of the study outlines the findings of the study which includes the data quality concerns which stem from the application of the PDQAF. Where the SDCAF is used to identify possible datasets, the PDQAF is used to examine the level of data quality exhibited by the particular dataset be it a survey or operational data register that is released. By applying the SDCAF fully, one is able to identify the particular questions or variables that are collected. This identification answers Question 3 of the study which requires the relevant attributes in datasets which describe the school dropout phenomenon to be identified. Furthermore, the findings which stem from the application of the PDQAF answer Question 4 of the study which requires the identification of impeding data quality concerns which limit the effective use of data to best describe the school dropout phenomenon. Chapter 9 of the study synthesizes the findings of the preceding 8 chapters of the study and outlines the learnings from the study, the recommendations which emanate from the examination of the school dropout phenomenon and data quality concerns and options for future research in these areas.

The last section attached to the document is the appendices which identify the relevant datasets, and applications of the PDQAF across each identified dataset from Brazil, India and South Africa.

	Chapter 1: Introduction
Part I: Theoretical Framework	• Chapter 2: School Dropouts in Brazil, India and South Africa • Chapter 3: Capability Approach, From Theory to Practice • Chapter 4: Data Quality, From Theory to Practice
Part II: Empirical Framework	• Chapter 5: Methodology • Chapter 6: Construction of the School Dropout Capability Assessment Framework • Chapter 7: Construction of the Public Data Quality Assessment Framework
Part III: Findings	• Chapter 8: Application and findings from the School Dropout Capability Assessment Framework and the Public Data Quality Assessment Framework • Chapter 9: Conclusion
Appendices	• Appendix 1 – Applicable data sources per functioning • Appendix 2 - Application of the Public Data Quality Assessment Framework for each identified survey • Appendix 3 - Data Quality Assessment of Brazilian Datasets • Appendix 4 - Data Quality Assessment of Indian Datasets • Appendix 5 - Data Quality Assessment of South African Datasets

Table 3: Chapter overview of the study

Part I

Theoretical Framework

Chapter 2

School Dropouts in Brazil, India and South Africa

Literature from Brazil, India, South Africa, and internationally conveys the multidimensional factors which leads a learner to choosing to drop out of school. As many authors argue, the choice is complex and is based on years of experiences that the learners accumulate during their school going career.

The key concern associated with dropping out of school, is the impact it has not only on the individual learner's opportunities in life but also the broader economic implications it has for a country when such a phenomenon becomes more widespread. Nicholas Spaull (2015) argues that those learners that do not complete schooling end up feeding mass unemployment or supplement the smaller proportion of unskilled workers in the country. These learners find themselves trapped in poverty with extremely limited means to rise into semi-skilled or formal labour. Byron Brown's (2010) qualitative study of a few learners corroborates Spaull's view that these learners were unable to find employment opportunities. In Brown's example, the learners attempted to re-enter the school system and were confronted with a diverse set of issues when attempting to do so. Spaull however, contends that it is usually the learners from low socio-economic circumstances that find themselves in the school dropout/poverty trap. The result of dropping out of school has never been an aspiration of learners, however as Spaull argues, it is the trap that these learners find themselves in.

If one accepts that a quantification of the school dropout phenomenon is more complex than the simple expression of the learner dropout rates as reported in each country, there is a consequent need to identify the factors which contribute to the individual learner's decision to stop attending school. The following section broadly sketches the needs and aspirations of learners from these three countries and amongst middle income countries in identifying reasons for their dropout. Various accounts from authors from these regions are explored in further detail.

2.1. Learner Dropout Factors in South Africa

The South African Department of Basic Education highlighted that the school dropout problem affects all grades in the country's schools. The dropout rate as defined by the department is the "proportion of pupils from a cohort enrolled in a given grade at a given school year who are no longer enrolled in the following school year" (Department of Basic Education 2011 p2) by following the United Nations Educational, Scientific and Cultural Organization's (UNESCO) 2009 Institute of Statistics guidelines in this respect. Using this definition and data from the National Income Dynamics Study (NIDS), in 2007/08, the average yearly drop-out rate between grades one to eleven amounted to 4% (Department of Basic Education 2011). The limitation of such a reported indicator is that it masks the inter-grade disparities and the cumulative attrition that the drop-out rate has on the education system as a whole. Nicholas Spaull (2015) contends the problem is much greater than the government recognises and emphasises that the throughput of learners is extremely low and as much as 50% of learners that start grade 1 actually enter grade 12. The differences between the reporting style of the department and academics such as Spaull, highlights the differences between an averaged aggregation and a cumulative system wide view of the phenomenon.

The differences in opinion regarding the problem are wide, with many who believe the department downplays the issue while shifting attention to issues such as the matriculation pass marks and avoids deeper issues that affect the education system. Various authors attribute many reasons for school dropout. The primary task of this study is to examine the links between such reasons and the underlying capabilities and functionings offered by the education system.

Martin Gustafsson (2011) attributes the quality of school education as the primary driver for high school dropouts in South Africa and low school completion rates. Gustafsson also recognises the important factors such as learner financial constraints and the provision of learner materials. Gustafsson highlights four factors such as

(1) the lack of household finance,
(2) the learner wanted to find employment,
(3) a drop in the learner's grades and
(4) pregnancy affecting female learners.

Karra and Lee (2012) reviewed trends of teenage pregnancy in the Cape area and their findings supported those made Gustafsson as they noted that females under the age of 20 in this area were 76% more likely to exit from school early and only 21% of such learners actually completed school by the age of 20. Spaull (2015) also highlights findings from the NIDS which strongly support the financial concerns raised by Gustafsson. These findings suggest that a key driver of school dropouts are learners that have difficulties 'keeping pace at school' and those that fall behind correlate strongly with their socio-economic circumstances of the learner and the school. Furthermore, simply repeating the grade is not sufficient to assist the learner in 'catching up'.

Fleisch, Shindler and Perry (2012) in their review of the Community Survey found that the key factors for dropout included learners dealing with a disability; whether the learner's family receives financial assistance in the form of social grants or if the parents of the learner were not employed. In addition, learners from households headed by children were found less likely to attend school. These learners were either the head of the household themselves who act as breadwinners or children under the care of the child household head. Fleish et al. also notes that the socio-economic conditions of the community that the learner resided in correlates with the learner's likely non-attendance of school. Branson, Hofmeyr and Lam (2014) provided a detailed review of the NIDS (Wave 2) in respect of school dropouts in South Africa. Whilst their study corresponded with Fleish's review of the Community Survey, the NIDS data analysis also highlighted an important finding that the quality of the school system had a stronger bearing on learner's proclivity to leave the school system. 'Falling behind' was found to be a stronger driver than simply the socio-economic status of the learner. Whilst socio-economic factors were important to consider, learners dropped out across income bands when they believed they were unlikely to succeed within the school system. Branson et al notes that 'falling behind' is a cumulative factor and therefore does not appear as the principal catalyst of leaving the school system. Branson et al's findings are one of few South African studies which correspond with international studies in this area.

Mnguni Bongani (2014) also highlighted the importance of recognising the household factors which impact the learner, as learners are more likely to dropout from school due to a poor guidance received from senior members of the household. Such learners are influenced by their peers and dropout from school due to a lack of personal appreciation for education and also due to the need to perform duties of the head of the household, such as act as a breadwinner for the household. Bongani also contends that learners act as the head of the household due to the effect of migrant parents who work and reside in another part of the country away from their children. In these instances, the parent infrequently visits home and the children are left to fend for themselves. Bongani also finds that the effects of discrimination and stigmatisation still occur within schools in Johannesburg. Often pregnant learners are discriminated against and chastised by learners and teachers alike. In some households, girls experience greater discrimination than boys do, as in these households, household members do not perceive that there is much value that a girl can derive from attending school.

The theme of child headed households was also found to have a significant impact in South Africa by Meintjies et al. (2009). The authors studied child headed households in greater depth and noted that the majority of child-headed households have at least a single parent that is alive, but is generally absent from the household. The eldest child is left to act as the head of household who thereafter becomes more prone to drop out from school. Such learners are more likely to live in informal settlements with poor access to basic services and many are located at great distances from the school and centres of business. In addition, their main sources of income are infrequently received remittances from family members or the household is a beneficiary to some social grant.

Two case studies from different provinces and municipalities in South Africa highlighted varied factors which predicted learner dropout. Flisher et al. (2010) tried to identify if there was a linkage between alcohol and drug abuse and dropping out from school. However, in the Cape Town schools that were surveyed, it was found that absenteeism, poverty and recent cigarette use were predictors of dropout, whilst drug and alcohol abuse did not have a direct bearing on learners leaving the system. Within the Fort Beaufort district, it was noted by Mgwangqa and Lawrence (2008) that the driving factor that predicted school dropout was extreme levels of poverty that the community had to endure. This study described poverty as inclusive of factors such as the unemployment of parents, hunger/starvation, insufficient income for school fee payment or school uniform purchase, walking long distances to school, lack of basic services at home as well as poor infrastructure and resources at the school. In addition to poverty, psychological consequences attributed to poverty were also apparent. This pertained to learner stress from parental neglect, general hardships, emotional problems, absentee parents and alcoholism. These studies again highlight the multitude of factors that correspond to poverty which tends to pull learners out of the schooling system.

Dieltiens and Meny-Gilbert (2012) examine the practice of social exclusion within schools and find that it’s occurrence is a common experience amongst poor learners who are excluded from accessing schools despite their right to education. In South Africa this practice is a consequence of the school attendance policies not being well understood by caregivers, learners and school administrators. Learners that are not able to pay for their school fees are victimised despite provisions within education policies which are meant to safeguard learners against such practices. Such learners are singled out during class-time and threatened with expulsion. Learners consequently feel intimidated and humiliated becoming progressively more likely to discontinue their schooling with each repeated incident. Learners that are conscious of their low-income status are stressed, anxious and feel inadequate due to their perceived lack of status amongst their peers. Interestingly, the authors highlight that these feelings of shame are most prevalent in communities with a greater degree of income inequality. Poor learners are found more likely to drop out from school if they come from poorer household compared to their peers as opposed to learners in schools who share a common level of poverty. Learners feel ashamed when they are unable to afford school fees, school uniforms and textbook amongst other resources required for school participation as compared to their peers. Learners thereafter feel inadequate and disempowered to live up to the perceived social and economic norm existing in the community.

Poor academic performance has been cited as one of multiple factors which in unison with socio-economic disparities faced by learners, collectively leads them out of school. A better understanding of the determinants of poor academic performance is useful to identify the root causes of the poor academic results and consequently school dropout. The following authors explore such determinants in South Africa.

Makgato and Mji (2006) in a study in South Africa identified a series of factors which influence the country’s poor academic performance especially in comparison to other middle-income countries. The first factor relates to poor teaching practices and the learning environment within the classroom. Makgato and Mji relate sentiments from learners which find that the teacher may not be patient enough especially in complex subject matters such as science and mathematics. However, interviewed teachers on the other hand note that learners are often not interested in learning. This issue of disinterest is a common factor also found in India. The second issue raised by Makgato and Mji involves how learners are taught complex subject matter. Learners raise concerns about being taught to memorise content instead of getting taught how to understand such content. Similarly, it was found that teachers feel inadequate to teach the more complex sections of the curriculum, and recommended greater training for teachers in such subject areas. The third factor is the poor laboratory facilities found in poorer schools. Complex subject matter may be more understandable if it can be practically experienced in a laboratory. However, if the laboratory is ill-equipped such experiments are impossible and learners are deprived of such opportunities for learning. The fourth factor was the learner’s non-completion of the syllabus in the lead up to the final examination for the year. Teachers and learners both suggested that there was not enough time in the calendar to complete the syllabus. The last factor identified was the indirect influence of parental involvement and language. Teachers expected parents to assist with homework completion, whilst learners often stated that their parents were ill-equipped to provide math or science support or they were too busy to be involved on a daily basis. Thus their home support for learning was insufficient in these cases. The issue of language related to the use of English definitions where learners that used English as a second language found it difficult to understand the definition’s meaning. Teachers also found that the vernacular translation could often be ambiguous and therefore avoided such translations.

Low teacher motivation was also identified as factor resulting in the poor academic performance of the learners. Lucille Petersen discusses the motivation of teachers and notes that in addition to the general competence of the teacher, their motivation for teaching had a direct bearing on the quality of their teaching (Petersen 2010 pp 44,45). Heystek and Terhoven (2015) unpack what drives educator motivation and the pair notes that the South African Basic Education Department motivation strategy involves acts of shaming the school and subject teachers by publishing pass rates of learners from the school. Supposedly poor performance amongst peers is expected to yield improved motivation improve performance standards each year. Where results are poor, educators are singled out and are threatened with losing their job. Where the results are positive, teachers are praised and attain a sense of accomplishment. Heystek and Terhoven argue that this approach has yielded few positive results.

In addition to the effects of poverty, teenage pregnancy and poor academic performance, factors affecting the mental well-being of learners were also recognised as a reason for leaving school. Topkin, Roman and Mwaba (2015) in their review of primary school educators’ knowledge of the disorder in South Africa found that they were only able to correctly identify less than 50% of the symptoms of ADHD, which therefore lead to a misdiagnosis of ‘acting out’ as simply bad behaviour in class instead of recognising the consequences of a neurological disorder. Topkin et al. argue that teachers need to be equipped with the skills to recognise the symptoms of ADHD and be enabled to institute interventions to effectively deal with the needs of their learners.

Stresses emanating from community factors also impact the learner’s mental well-being. This has been linked to not feeling safe at school, which has also been found to pull learners out of school. Burton and Leoschut (2013) report that at least 22% of learners in public high schools were affected by violence in 2012 in the form of assault, robbery, sexual assault and cyber violence. Furthermore, learners that are repeatedly affected believe the likelihood such occurrences declining is slim. Violence not only affects the learner but the educator as well. The authors report that educators are also threatened and attacked verbally and physically by learners who they teach. Burton and Leoschut cite access to alcohol, drugs and weapons as reasons for such behaviour. Furthermore, learners are influenced by gangsterism and other ill-effects from within their communities which are expressed within social environments in communities such as schools. These cases lead to climate of fear within the school which affects learners during class time and on their journey to and from school.

Sabates et al. (2010) who conducted a study with the University of Sussex, noted that in many low to medium-income countries, including South Africa, although there are greater numbers entering the school system, higher percentages of learners are not completing primary school. The reasons cited for these dropouts include limited learning opportunities, overcrowded classrooms, children of different ages and abilities mixed in a single class, teaching methods that are not adapted to the different ranges of learners included in a class. In addition, there were family-level factors such as ill-health, malnutrition and poverty. Sabates et al, emphasise the multi-dimensional nature of factors that drive school drop-outs in these countries and further highlight how similar factors affect a diverse range of countries in the middle-income band. They find it is the combination and cumulative effect of attending schools affected by such issues which together tend to influence young learners to leave the school system. For the purposes of this study it is important to identify whether similar factors apply to India and Brazil. In summary of the factors identified in the preceding section, Table 4 presents the key issues that each author identified.

Author	**Reason for dropping out**
Gustafsson	• The lack of financing, • The learner wanted to find employment, • A drop in the learner’s grades • Pregnancy amongst female learners
Spaull	• Socio Economic Context of the learner • Quality of learning
Fleisch et al.	• Poverty of the Household • Disability status • Child headed households • Socio-economic status of the community
Mnguni Bongani	• Lack of guidance or poor guidance received from senior members of the household

Author	**Reason for dropping out**
	• Child Headed Households • Migrant workers residing in another part of the country away from their children • Discrimination against pregnant learners • Opinions in household or community that girls do not need to attend school
Lucille Petersen /	• Low teacher motivation
Meintjies et al	• Migrant heads of household • Infrequently received remittances from migrant breadwinners, force learners to seek employment
Branson, Hofmeyr and Lam	• Quality of the education system • Socio-economic factors
Sabates et al.	• Limited learning opportunities • Overcrowded classrooms • Children of different ages and abilities mixed in a single class • Teaching methods are not adapted to the different ranges of learners included in a class • Learner's ill-health and malnutrition • Poverty and socio-economic status
Flisher et al.	• Absenteeism • Poverty • Recent cigarette use
Mgwangqa and Lawrence	• Poverty ○ Unemployment of parents, ○ Hunger/starvation, ○ Insufficient income for school fee payment or school uniform purchase, ○ Walking long distances to school, ○ Lack of basic services at home ○ Poor infrastructure and resources at the school • Psychological Factors Linked to Poverty ○ Parental neglect ○ General hardships ○ Emotional problems ○ Absentee parents ○ Alcoholism
Dieltiens and Meny-Gilbert	• Shame of poverty leads to social exclusion
Makgato and Mji	• Poor academic performance • Poor teaching practices • Unconducive learning environment • Poor or ill-equipped laboratory facilities • Ill prepared teachers in complex subject matter • Non-completion of the syllabus leading up to the final examination • Parents are ill-equipped to provide math or science support • English is a barrier when learning definitions of complex subject matter.

Author	Reason for dropping out
Topkin, Roman and Mwaba	• Impact of learning disorders on the student • Misdiagnosis of ADHD or other disorders by the educator
Burton and Leoshut	• Crime in the community • Crime at school • Feeling unsafe at school

Table 4: Summary of learner dropout factors in South Africa

2.2. Learner Dropout Factors in India

The Ministry of Human Resource Development in India in the CABE Report on the Universalisation of Secondary Education noted that in 2004 the ministry it was working towards a zero dropout rate. According to the CABE report grades eight, nine and ten experienced a 10% dropout rate; however, the greatest difficulties were experienced at the final board exam in which 50% of learners failed. Therefore, the combination of dropouts and low pass rate at the final leg of the school system highlight the limited number of learners who successfully negotiate the path through the school system. Similarly to South Africa, the quality of school system is recognised as a primary concern limiting the throughput of learners through the education system (Central Advisory Board of Education 2004).

Sunita Chugh (2011) stated that learner dropouts are a widespread problem affecting India. In India, gender is a more influential factor driving school dropout. Chugh highlights in her study that 57% more girls than boys dropped out from school. It was noted that families were concerned for their daughter's safety, often fearing the possibility of rape in the slums. This fear combined with a belief that female learners needed to rather contribute to the maintenance of the household resulted in higher dropouts among female learners in India. In addition to gender disparities, socially disadvantaged castes and tribes tended to have higher percentages of learners that dropped out from school. The study also finds factors related to the learner's family in addition to their school which influences their choice to drop out of school. Therefore, as Chugh suggests, it is not only poverty that is a factor but there are also broader issues in the school system which are not responsive to specific education needs of the learners. In relation to poverty, the issues identified were the cost of school fees and the various resources required in school attendance such as books and uniforms. Family issues referred to the disputes that households experienced such as financial constraints or the effects of excessive alcohol consumption. In terms of school issues raised by Chugh, it was found that schools in need of repair or lacking facilities, together with the learner's need travel long distances to attend school, were found to gradually pull learners out of school. The final issue raised by Chugh were child specific issues that dealt with the learner's ability to grasp elementary level concepts or those who performed academically weaker in comparison to their peers. Generally, those that struggled academically would often drop out of school.

Sajjad et al. (2012) studied schools from South East Delhi and found that school dropouts were also a pervasive problem across India amongst primary schools. Poverty was once again found to be a driving factor that excluded learners. Sajjad noted that when families were incapable of providing for their basic subsistence, the financial demands of schools were deprioritised in favour of accommodation or food expenses. The main factors that the study highlighted were the disparities between the genders and the influence of factors such as family type, lack of income, unemployed heads of households as well as the education level of the household head. Sajjad et al's research referred to various other pieces of studies which corresponded with their study but also highlighted additional factors such as larger family sizes, caste disparities and poor infrastructure development of the school.

Gouda and Sekher (2014) in their review of Indian literature and the National Family Health Survey-3 found many of the factors highlighted by Sajjad in Delhi corresponded with the national view. Again it was found that those individuals in lower castes were more likely to leave school. In addition, it was also found that Muslim learners were also more likely to dropout from school. The other major factors identified related to the characteristics of the parents of learners. Learners whose parents were illiterate or not working were more likely to drop out. The main reasons identified directly with the National Family Health Survey were (1) learners not interested in further study, (2) excessive costs of fees, (3) learners were required to work and (4) the learner had failed repeatedly up to that point.

The various forms of discrimination which affect learners have also been found to be a factor that pulls learners out of school. Dubey (2010) discusses the rampant discrimination between the rich and poor. Whilst the rich have access to high quality, private education, the poor (inclusive of minority population groups and the marginalised) only have access to government schools which are often in a dilapidated state. Dubey believes India should implement norms and standards to govern the implementation of anti-discrimination policies. Hussain and Chatterjee (2009) discuss examples of discrimination that female learners face, such as, teenage girls who are not consulted about their early marriage. In addition, there is also rife discrimination amongst religions in rural communities especially Muslims, for example, who are often mistreated. An additional dynamic pertains to the combination of religion and gender where female Muslim learners are more likely to complete school than female Hindu learners due to the differences in cultural beliefs regarding the role of females in the household. Srivastava (2014) echoed the concerns raised that affect female learners in India and suggests that the curriculum of India needs to include greater gender sensitivity training in order to inculcate a culture where female learners and their respective households can begin to value female school completion to a greater degree.

Kumar, Panda and Jena (2013) conducted a detailed statistical analysis of a survey of youth who had dropped out of the school system. Various questions were posed to the former learners regarding their decision to leave school. The study sought to identify the associative links between the questions in the survey, in order to identify the most important factors which affected the learners. Four key groupings of factors were identified which tended to agree with the previously mentioned factors. This included poverty, the teaching environment and the disinterest in continuing to study. The teaching environment related to the infrastructural level of the school as mentioned previously and also included the poor quality of teaching received. The disinterestedness factors were linked to learners suggesting they had weak memories or a fear of study. Their parents influence over their studies also affected their level of disinterest.

Frances Hunt's (2008 pp 7-17) review of South Asia and Sub-Saharan Africa provides a unique perspective on the commonalities of these areas and he finds various factors at a macro-level that affect both areas. Hunt grouped the dropout factors he identified in terms of household finances, contextual issues emerging from the household, health in general and the social and political context found in the country. The issues related to the household income highlight the challenges of poverty which affect many households in India. Where households are affected by a sudden loss of income due to job loss, learners are asked to sacrifice their schooling as it is considered to be a luxury in the context of household expenses. Often the child is also required to supplement the household income as described by Gouda and Sekher. In terms of the contextual issues related to the household, shocks to the household such as family bereavements accompanied with a loss of income results in the learner carrying some of the financial burden experienced within the household. Together with household poverty, it was found that parents in poor households often holds a negative perception about the value of school on their child's life. Such parents have often not completed schooling themselves and pass on these points of view to their children.

In terms of the health issues raised by Hunt, poor households tend to prioritise medical care costs over the need to attend school. Furthermore, poor households often suffer from malnourishment which impedes their concentration, reduces their cognitive abilities in class and results in low motivation to attend school. Within the health domain are factors such as teenage pregnancy and physical impairments of learners. Teenage pregnancy is also a concern linked to factors of poverty and poor learner academic performance. Disabled learners and those with special needs often require greater financial support and require their schools to be equipped to meet their requirements such as wheelchair ramps or textbooks provided in braille. Such facilities are often neglected as they are costly for both the poor household and schools found in rural and poor areas (Hunt 2008 pp 24-30).

In terms of the social and political context, in some communities, Hunt found a bias against female learners' learning prospects. In such areas, societal values that encourage the males to lead households whilst the females are prepared for marriage and are influenced to discontinue their education. Hunt also found that female learners were kept away from school when they started to reach maturity in an effort to save their reputation of purity which was considered necessary for marriage. In these cases, communities valued marriage of females more than school completion (Hunt 2008 pp 30-37).

The rural-urban divide is also a key factor in India. Learners in rural areas tend not to complete schooling but often leave to take up work in support of the household. In addition, in some rural areas, the final secondary schooling grades are not offered in the public schools in such areas without any alternative in the area. To complete schooling, learners are forced to travel great distances. Within India, Hunt also identifies socially repressed communities based on caste, religion and ethnicity. Such communities suffer from discrimination in the form of verbal abuse from teachers and classmates or for example, being tasked to perform humiliating jobs whilst at school (Hunt 2008 p33). Furthermore, the influences of conflict or violence within the community make school attendance difficult. Such factors when viewed from a macro perspective highlight how violence and conflict restricts access to school not only for the learner but the teachers and support staff as well.

Hunt's final social and political factor involves when a child enters the school system. The late entry of a learner to school results in their late completion of schooling. In such situations, the learner, when ready for senior levels of secondary school, is already considered to have reached adulthood in the community and must therefore share the burdens of household management and support (Hunt 2008 pp 44, 45). This notion of seniority in the household is contrasted with the perceived treatment of being a school child and tends to lead to greater absenteeism and ultimately school abandonment.

Henal Shah (2005) examined the feelings of shame that learners experience arising from academic failure. The learner’s successful completion of school examinations is seen as the learner’s gateway to a successful life whilst the stress of failure of these examinations results in psychiatric difficulties for the learner. Shah noted that poorer learners experience greater levels of such stress and tended to dropout out of school. In addition to dropping out of school, learners develop temperamental difficulties which impedes their academic performance. Furthermore, stress from parental pressure, peer pressure, school pressure and external pressure all add to a poor performing learner’s loss of self-esteem, anxiety and suicidal tendencies in extreme cases. Shah notes that in a study in 1997, up to 20% of Indian learners experienced such emotional struggles.

Neelam Sood (2010) discusses the complex problems of environmental deprivation factors such as poverty, poor health-care and weak support structures in the family and community. Whilst Sood argues these factors are applicable internationally, her study focusses on the Indian context. Sood also believes that malnutrition has a direct bearing on the learner’s willingness to participate in school and their resultant dropout, even though literature rarely makes the connection between the importance of nutrition and school enrolment. Malnourished learners tend to be more prone to disease and such learners are predisposed to cognitive underdevelopment affecting their attention span and ability to concentrate in school. These factors result in higher dropout rates. Similarly, a study by Banerjee, Das, Shinkre and Patel (2011) of schools in the Goa area noted how hunger had a direct impact on academic performance, physical health and mental well-being.

From this cursory review of studies from different perspectives one finds that the factors for school dropouts remain fairly consistent. They often refer to the socio-economic conditions of the learner’s household, community and school and also refer to issues of caste, religion and gender. These factors are summarised in Table 5 below. Following this enumeration of factors, the relevant factors in Brazil are explored in the next section.

Author	**Reason for dropping out**
Chugh	• Gender ○ Insecurity of girls attending schools • Caste and Tribe • Poverty ○ Cost of fees, private tuition, resources such as stationery, uniform, transport • Home Factors ○ Disputes within a family related to job loss, financial constraints ○ Alcoholism

Author	**Reason for dropping out**
	• School Issues o Attitudes of teachers in poor schools o Infrastructural facilities o Distance of School to the home • Child specific issues o Poor understanding of elementary level work o Poor academic performance o Early Marriage o Education levels of the parents o Employment level of the parents o Insufficient space to study at home o Illness of the child or family member
Gouda and Sekher	• Caste, Muslim and Girl learners • Education levels of the parents • Employment level of the parents • Not interested in studying • Cost of fees • Required to work in the household or to supplement household income • Repeated failure
Dubey/ Hussain and Chatterjee/ Srivastava/Raju and Singh	• Discrimination between rich and poor • Discriminations amongst castes and religions • Female learners forced to marry young
Kumar, Panda and Jena	• Poverty o Unemployment of parents, o Hunger/starvation, o Insufficient income for school fee payment or school uniform purchase, o Walking long distances to school, o Lack of basic services at home • Teaching Environment o Infrastructural level of the school o Poor teaching quality • Disinterestedness o Weak memory o Fear of study o Impact of Parent
Hunt	• Household Finances o School Fees, o Income shocks to the household o Child Employment o Migration • Household context o Family bereavement o Education level of household members o Perceived benefit of schooling • Health o Health of children and relatives (Malnutrition, hunger, illness)

Author	Reason for dropping out
	○ Pregnancy ○ Disability and special education needs • Social and Political Contexts ○ Gender ○ Rural/Urban divide ○ Socially disadvantaged grouped ○ Conflict, Emergency situations ○ Age, marriage and notions of adulthood
Sood/Banerjee	• Poverty • Poor health-care • Weak support structures in the family and community • Malnutrition of learners • Impact of school nutrition programmes
Henal Shah	• Feelings of shame learners experience arising from academic failure • Stress from parental pressure, peer pressure, school pressure and external pressure

Table 5: Learner Dropout Factors in India

2.3. Learner Dropout Factors in Brazil

The World Bank has noted that Brazil’s education system has been greatly revitalised over the period of 1995 – 2010. Most education indicators in Brazil in 1994 represented various states of precariousness across the system. In 1990 the average years of schooling per learner amounted to 3.8 years and this figure had risen to 7.2 years by 2010. In addition to the improvement in school retention and completion, the quality of the education system has also been noted to have improved and Brazil was identified as having “increased educational attainment of the labor force faster than any other developing country, including China, which had set the global record for schooling expansion in the prior decades” (Human Development Sector Management Unit 2010 p14).

Brazil’s education policies over this period have been credited with the country’s remarkable progress. It has been noted that Brazil’s national education policies follow global best practices and focus on education finance equalisation. Brazilian policies therefore focused on funding the disparities between learners as well as closely monitoring the results of the programmes with a strong emphasis on data management. This required Brazil building a world class education measurement system, secondly closely incorporating it into policy design and thirdly, utilising conditional cash transfers directly to the poor using innovative channels of providing such funding. The conditions attached to the funding required low-income households to ensure their children attend school (Human Development Sector Management Unit 2010). In this light, due to the improved socio-economic conditions in Brazil, it is likely that factors such as poverty may not be attributed as greatly as found in India and South Africa as a reason for exiting the school system.

Soares et al. (2015) analysed a survey of the Minas Gerais in Brazil to identify predictors of early school exit by learners. The study highlighted the significance of factors such as difficulties learners experience in understanding their subject content, the learners wanting to change schools, the learner's perception that school completion equates with improved employment, and the learner's preference to attend a particular chosen school.

Cardoso and Verner (2006) analysed the urban poor regions in Fortaleza Brazil in an attempt to identify the factors which cause school dropout. Cardoso and Verner sought to test whether child employment or pregnancy excluded learners from the school system. The intention was to determine whether the conditional grants from the state were effective mechanisms in maintaining school attendance. Due to the grant provision, it was found that the need for employment was not a driving factor pushing learners out of school. Interestingly, in Brazil, more boys dropped out of school than girls, but that this fact did not reduce the impact that pregnancy had on the rate of school dropout. Furthermore, it was noted that extreme poverty and especially hunger was a factor for learners dropping out of school. Cardoso and Verner also discuss the traditional factors that are generally identified in Brazilian literature and these include drug use, alcohol abuse, parents with psychiatric disorders, socioeconomic conditions and general demographics such as gender, race and age.

Graeff-Martins et al. (2006) studied schools in a city in Brazil and found that a key factor in these schools was absenteeism coupled with learner's experiencing mental health challenges. The mental health disorders were found to more likely to cause learners to drop out. Such mental health issues included attention-deficit disorders, depression, anxiety, mental retardation and substance abuse.

Tramontina et al. (2001) linked the school dropout to the high rate of legal and family problems that often beset Brazilian society. The key factors identified by Tramontina included age, gender and ethnicity discrimination, learner with a lower comparative estimated IQ, parents who did complete school, grade repetition amongst learners, single parent families and households faced with low monthly family income. Each of the factors have been found common in other studies in both South Africa and India. These factors from each of the above mentioned authors in Brazil are summarised in Table 6 below.

Author	Reason for dropping out
Soares et al	• Socio Economic Index • Gender • Pregnancy • Need to Work to supplement family income • Age-Grade Gap • Difficulties with subjects • Desire for a different school • Perception of better job opportunities when completing school • Importance of school choice
Cardoso and Verner	• The role of early parenthood • Extreme Poverty + hunger • drug use, • alcohol abuse, • parents with psychiatric disorders, • Socio-economic conditions, • gender, race and age
Graeff-Martins et all	• Absenteeism • Learner's state of Mental Health (attention-deficit disorders, depression, anxiety, mental retardation and substance abuse)
Tramontina et al	• Age • Gender • Ethnicity • School Repetition • Estimated IQ • Parents educational level • Single Parent Families • Monthly family incomes

Table 6: Learner Dropout Factors in Brazil

2.4. Internationally Recognised Learner Dropout Factors

Many factors have been identified internationally which impact middle-income countries or have been raised by international bodies which highlights the universalism of such factors. Relevant factors which impact Brazil, India and South Africa that were not directly raised by authors from these countries have been listed here.

The World Food Programme (2010), notes that the adoption of school feeding programmes can improve micronutrient and macronutrient intakes which promotes child health, increases the effectiveness of learning and decreases morbidity amongst poor learners. Such programmes are also enticing to learners from poor backgrounds and therefore promotes school enrolment and attendance in school.

There are studies in Sub-Saharan Africa which point to female sanitary-care which is available to poor, pre and post pubescent female learners. Jewitt and Ryley (2014) note that menstruation is a critical factor that influences large increases in female school absenteeism and female school dropouts and drives gender inequalities. These topics are generally considered taboo and therefore there is little empirical evidence supporting the notion. However, the few studies conducted in Kenya highlight how poverty exacerbates gendered bodily inequalities. Jewitt and Ryley also contend that the poor school-based sanitation programmes is a strong underlying reason for poor female school attendance as schools that provide access to sanitary products have higher female enrolment. It was noted that there is a clear gender bias amongst school dropouts, especially in developing countries where female learners leave school due to puberty. Female learners are more inclined to leave school after regularly missing school due to the school not having adequate sanitation facilities to accommodate their periodic menstruation.

Kirk and Sommer (2006) conducted a study across Sub-Saharan Africa and Asia and linked the girl's health to the girl's education. The authors concur with the issues raised by Jewitt and Ryley and note that the poor provision of sanitary supplies can impede a girl's school access and experience. The authors therefore argue for school-based programmes to address menstrual or female maturation related concerns. Kirk and Sommer note the relevance for such a programme in India where menstruating girls are considered untouchable in some communities and are shunned. The author attributes this to a lack of knowledge about the natural biological process that female learners undergo. The lack of factual information compounds the spread of myths which ultimately victimises poor female learners. Kirk and Sommer also identified the cost of sanitary products as a major factor. It was found that the monthly cost of sanitary protection equated to the monthly cost of batteries for a radio or a household's monthly paraffin consumption. Poor households are faced with such trade-offs when budgeting for their needs and female sanitary needs are often deprioritised. An additional health factor which accompanies menstruation is the simple physical discomfort which is often an unattended issue faced by poorer learners in developing countries. Discomforts such as lower-back pain, bloating and cramping can be attended to by a range of pharmacological products such as pain-killers. However, such products are simply inaccessible to poor households and are viewed as luxury items.

Grant and Hallman (2008) also refer to other studies that discuss how poorer learners, who are not academically strong, are more predisposed to engage in unprotected premarital sex. The arguments for why learners that are academically struggling, engage in such risky behaviour is largely anecdotal but points to the learner's individual value system. Such learners may not see any real prospects in school and this notion combined with a low understanding of the risks of unprotected sex leads to an often unintended consequence of early pregnancy. Grant and Hallman note the difficulty in obtaining data of pregnant school girls and rather argue that one should attempt to identify the factors which lead to early teenage pregnancy which may provide policy makers the ability to address the trend. The factors identified by the authors relate to a gender biased community, the household's financial strength, late school entry, lack of access to contraceptives and lack of education regarding the risks of sexual behaviour. To combat the effects of pregnancy, the authors believe pregnancy prevention programmes should be promoted which provide training to adolescents both male and female, of the risks of pregnancy whilst also making contraception more accessible.

Thurlow, Sinclair and Johnson (2002) note that learners that suffer from learning difficulties are much more likely to repeat a grade or dropout from school. Statistics from the United States highlight that 36% of learners that dropout from school suffer from a learning disability whilst 59% of learners contend with emotional or behavioural disabilities. Lund et al. (2011) argue that mental ill health and poverty work together in a continuous negative cycle as the longer a person is found to live in poverty, the greater their risk of mental illness through exposure to stress, social exclusion, malnutrition, violence and trauma amongst other factors. The converse argument that applies is that the likelihood of drifting into poverty is greater amongst those that suffer from mental illness due to an incapability to be a productive employee linked with the related loss of income. Among the neurodevelopment disorders which affect learners are the specific learning disabilities which affects the learner's ability to read (dyslexia), to write (dysgraphia) and to conduct simple mathematic calculations (dyscalculia). These disorders do not limit the learner's cognitive ability to comprehend a particular lesson, but rather due to their limitations to read, or write or perform math find that they generally require more time to complete a particular class exercise (Karande 2008).

Aro et al. (2011) discuss how the lack of recognition of the learners' learning deficiencies tends to impact on the learners' confidence in their cognitive abilities. A recurring failure due to difficulties experienced with the basic learning activities such as reading, writing and math discourages a learner to continue school participation and

Porche et al. (2011) note that learners from poorer communities that must contend with continued exposure to stress, crime, trauma and violence experience the greatest impact of such effects and this is presented in their levels of academic achievement, greater absenteeism and greater dropout rates. Porche et al. argue that a learner's experience of a traumatic event or chronic exposure to a highly stressful environment may lead to psychiatric disorders. Trauma can be experienced in the form of physical and sexual abuse, domestic violence, community violence, school violence, bullying, severe neglect, traumatic injury or the traumatic loss of a loved one. Porche et al. further argue that the sustained experiences of stress, adversity and trauma disrupt the brain architecture and the learner's body develops an internal stress management system which ultimately reduces the learner's cognitive information processing ability. The authors argue that poorer communities require a greater integration of school staff and mental health professionals that can assist with the recognition of such occurrences and are thereafter able to recommend particular courses of action.

Robert Blum (2007) in a study for the United States Military attempted to identify the best practices that ought to be employed in setting up a school and enhancing the school environment. Blum believes the school environment refers to the social, academic and emotional contexts that the school provides to deliver and enable learning. The learners, staff and community form perceptions of the school within these contexts. It is important therefore that the school is able to provide resources that enable learners to form a positive opinion of the school in terms of these contexts. Blum believes the first priority for a school is to provide a safe and structured environment. In this regard, Blum argues that safety is a central tenet for schools to address which entails ensuring that school resources are protected and that learners feel safe.

Porteus et al. (2000) and various other authors have expressed the view that the distance from school is a central school-related factor for learners dropping out of school. Whilst traveling great distances is a factor, the associated costs of transport to school are also a key factor to consider for poor households. Chinyoka (2014) also notes that together with traveling great distances, we find that learners are exhausted from the travel before their actual school day begins. Furthermore, after traveling such distances, learners often arrive late to school and are consequently admonished for tardiness and marked as absent in school attendance registers. Such experiences demotivate school learners whom regularly face such challenges. An additional factor to consider regarding why learners need to travel great distances is due to the isolation of rural communities. Smink and Reimer (2015) note that rural areas have limited transportation opportunities compared to those in urban areas. Shahidul and Karim (2015) notes the vulnerability female learners feel when traveling long distances to school due to sexual harassment. As a consequence of these feelings, female learners have been found to be more likely to drop out of school.

Richard Roe (1991) notes that a learner’s free expression of their beliefs and views follows from Constitutional protections; however, school authorities tend to restrict free speech in the aims of maintaining discipline and orderliness in a school. Pedagogical research highlights the learner’s ability to freely express themselves fosters their cognitive development and their appreciation for school enrolment. De Waal, Mestry and Russo (2011) discusses these limitations on learners and highlights how such infringements on a learner’s rights tend to push the learner out of school as they feel persecuted for expressing their religious beliefs. Often these are minorities in the community who bear the brunt of such rulings as their attire differs from the identified community norm.

These factors from international studies are summarised in Table 7 below. The final section of this chapter follows thereafter with a presentation of the synthesis of all factors presented thus far.

Author	**Reason for dropping out**
World Food Programme	• Access to school feeding programmes • Micronutrient and macronutrient intakes of learners
Jewitt and Ryley/Beksinska et al	• Menstruation • Poor school-based sanitation programmes • Costs of sanitary products for female learners • Accessible sanitation facilities at school
Kirk and Sommer	• Poor provision of sanitary supplies for female learners • School-based programmes to address menstrual or female maturation related concerns • The monthly cost of sanitary protection
Grant and Hallman	• A low understanding of the risks of unprotected sex • Early teenage pregnancy
Thurlow, Sinclair and Johnson/Lund et al./Karande/Aro et al./	• Learning disabilities (dyslexia, dysgraphia, dyscalculia) • Mental ill health
Porche et al	• Exposure to stress, crime, trauma and violence • Physical and sexual abuse • Domestic violence, community violence, school violence, • Bullying • Severe neglect • Traumatic injury • Traumatic loss of a loved one • Integration of mental health experts
Blum	• Resources that enable learners to form a positive opinion of the school in terms of Social, academic and emotional contexts • Schools with low security

Author	Reason for dropping out
Porteus et al./Chinyoka/Smink and Reimer/Shahidul and Karim	• Distance from school • Limited transport services in rural areas • Sexual harassment of female learners that travel long distances to school
Roe/De Waal et al.	• Learner’s ability to freely express themselves

Table 7: Learner Dropout Factors raised internationally

2.5. Synthesis of Learner Dropout Factors

On review of the above factors across South Africa, India, Brazil and internationally, various commonalities exist. The countries share many challenges whilst some issues that are highlighted within the identified studies are more specific to a particular country. When these factors are grouped by the granularity of the context, the below categorisation emerges as shown in Table 8. Many of the authors focus on home, school, child and various demographics, whilst many of the factors were applicable to the context of the community where the learners reside. Often learners of a particular community were more likely to leave the school system due the living conditions of that particular area. Therefore, on review, the analysis of the factors is presented in decreasing levels of granularity, viz., Community, School, Household and Individual, with an inclusion of relevant demographic categorisations.

Context	Factor	South Africa	India	Brazil	International
Community	Socio Economic Conditions	Yes	Yes	Yes	Yes
	State of Poverty	Yes	Yes	Yes	Yes
	Experiences of crime/violence				Yes
	Level of Unemployment	Yes	Yes	Yes	
School	Infrastructural State of School	Yes	Yes	Yes	Yes
	Provision of security at school				Yes
	Cost of Fees	Yes	Yes	Yes	
	Access to Resources	Yes	Yes	Yes	Yes
	Access to Sanitation				Yes
	Distance to School	Yes	Yes		Yes
	Overcrowded Classrooms	Yes			
	Absenteeism	Yes	Yes		Yes
	Negative Teacher Attitude, Motivation		Yes		Yes
Household	Household Income	Yes	Yes	Yes	
	Access to Basic Services	Yes	Yes	Yes	
	Child Headed Household	Yes			

Context	Factor	South Africa	India	Brazil	International
	Employment Status of the Parents or Head of Household		Yes	Yes	
	Education Levels of the Parents or Head of Household		Yes	Yes	
	Parents with psychiatric disorders			Yes	
	Family Disputes/Violence	Yes			Yes
	Alcohol Abuse	Yes	Yes	Yes	
Individual	Estimated IQ			Yes	
	Psychological Factors	Yes	Yes	Yes	
	Trauma, Stress, Abuse				Yes
	Learning disabilities				Yes
	Disinterest in Learning		Yes		
	Grade Repetition	Yes	Yes	Yes	
	Hunger/Starvation/Malnutrition	Yes	Yes	Yes	Yes
	Absenteeism	Yes	Yes		
	Poor understanding of elementary level work		Yes		
	Poor academic performance	Yes	Yes	Yes	
	Early Marriage		Yes		
	Teenage Pregnancy	Yes	Yes	Yes	Yes
	Knowledge about unprotected sex				Yes
	Desire for a different school			Yes	
	Importance of School Choice			Yes	
	Gender	Yes	Yes	Yes	
	Menstruation, provision of sanitary supplies, Costs of sanitary supplies				Yes
	Freedom of expression				Yes
	Learner motivation				Yes
Demographics	Age	Yes	Yes	Yes	
	Race	Yes		Yes	
	Caste		Yes		
	Ethnicity		Yes		
	Religion		Yes		
	Disability	Yes			

Table 8: Synthesis of Learner Dropout Factors

Community related factors included issues such as the general socio-economic conditions that affected the learner, the state of poverty and the level of unemployment. These factors were consistently raised amongst all three countries.

At a school level if one considers the infrastructural state of the school, the enrolment fees, the ranges of resources available to a learner in a school, the distance a learner must travel to attend school, overcrowded schools, absenteeism and a negative teaching attitude. The distance to school issue was not raised amongst Brazilian studies and neither was absenteeism or a negative teacher attitude, whilst overcrowded classrooms were only raised in South African studies. The issue of negative teacher attitudes also wasn't raised in South Africa. Internationally, the access a female learner has to sanitation facilities influenced their absenteeism due to effects of menstruation.

Amongst households the income levels, their access to basic services, whether the household was headed by a child, whether a member of the household was employed, the education status of the head of the household, any psychiatric health issues of the learner's parents, internal family disputes and alcohol abuse were identified as concerning issues affecting each country. Child headed households seemed only pertinent in South Africa whilst South Africa did not seem affected by the employment status and education levels of parents of the learner. Psychiatric disorders were raised only within Brazil whilst family disputes were only highlighted in South Africa.

At an individual level, the following factors were considered very important, viz., the estimated IQ of the learner, psychological factors affecting the child's learning, learning disabilities, a disinterest in learning, frequent grade repetition, general hunger or malnutrition, regular absenteeism of the child, a poor understanding of elementary level work, poor academic performance, early marriage, teenage pregnancy, the desire to change school and the importance of school choice. The estimated IQ of the learner was a notable factor in Brazil whilst the disinterest in learning, the poor understanding of elementary level work and early marriage were only raised in Indian studies. The desire for a different school and the importance of the school that was attended was only notable in Brazilian studies. Internationally, not only was pregnancy noted but also the knowledge that learners had about the dangers of unprotected sex was also recognised as a factor.

In terms of demographics, age, race, caste, ethnicity, religion and disability were highlighted. Race was not relevant with India, whilst the country faced challenges such as caste, ethnicity and religion that were more specific to their context. Issues of disability were only raised within the reviewed South African studies.

Considering the wide range of factors identified at varied levels of disaggregation, one notes that sourcing data to describe such elements can be difficult. Either the data has not been collected, has limited exposure or there are a range of quality issues affecting which sets of such data are usable. In addition to assessing the quality of the data, an exercise is still required to identify how Sen’s Capability Approach can be best utilised to identify the most pertinent data sets that describe the learner’s aspirations in terms of core themes which describe school dropouts. Therefore, it will be necessary to develop a framework that will help assess these themes in terms of the Capability Approach and the levels of data quality we find in available data sources.

Chapter 3

The Capability Approach: From Theory to Practice

This chapter explores the development of the Capability Approach Theory and discusses the significance of using a non-commodity based measure to describe the pluralist/multidimensional view of human development. The Capability Approach as described by Sen offers the theoretical grounding to identify the level of real freedom that a person, rich or poor, can attain contrasted against a collection of aspirations (described as functionings) that the individual finds most important in life. Each person's valued functionings differs due to the different opportunities available to a person and therefore their state of real freedom varies as well. Furthermore, the selection of an array of functionings which describe the state of one's attainment of valued goals in life, is empirical in nature and lends itself to the enumeration of indicators which describe the multidimensional nature of freedoms that one wishes to attain. This is discussed and theorised across the expanse of Amartya Sen's (1992, 1999, 2000, 2009) body of work in formulating the Capability Approach Theory.

The recognition of this empirical quality, drives the proliferation of applications of the approach, together with authors finding it useful to describe the pluralistic nature of complex phenomena within their country of study. Through the work of United Nations Development Programme (UNDP) (1990), Ul-Haq (1995), Nussbaum (1993, 2000, 2011), Alkire (2002a, 2002b, 2007, 2008, 2009) and others, lessons have been learnt which assist in defining the elements related to a quality of life that individuals wish to attain and the principles that should be employed when attempting to gather data which describe a particular complex multidimensional phenomenon.

Within the Education sector, the work of Unterhalter (2009, 2015), Walker (2005a, 2005b) and Saito (2003), amongst others, provide concrete examples of how the approach has benefited analysis in the sector. However, one also recognises the gaps that are introduced in a study of learners if the focus on factors affecting the learner is limited to the school environment alone and not to the wider household, community, political and economic factors as well. To this end, a broad human development focus is needed to understand the wide range of themes which impact the learner. Thus the work of the UNDP (1990) Ul-Haq (1995), Nussbaum (1993, 2000), Narayan et al. (2000) and others provide the foundation for formulating a Capability Set which can be used to describe the factors which best describe school dropouts. This structure can thereafter be supplemented with the school dropout literature from the three countries as discussed in Chapter 2.

From the analysis of the various Capability Approach studies, one is able to define the central real freedom which is assessed in this study which is noted as the learner’s real freedom to attend school in preparation for attaining employment and a better quality of life. To assess the attainment of this real freedom across Brazil, India and South Africa, data is collected to describe each functioning within the identified Capability Set. This includes the learner’s financial security, physical and mental well-being, the state of their school infrastructure, the factors related to school travel, a conducive learning environment, effective school participation and one’s ability to freely express themselves. The formulation of this Capability Set, lays the foundation for answering research Question 2 of this study which concerns which framework can be adopted to describe the School Dropout Phenomenon.

3.1. Capability Approach Theory

3.1.1. Foundation of the Capability Approach Theory

The Capability Approach as developed by Amartya Sen (2000, p19) and discussed in his work “Development as Freedom” differs from other policy analysis and evaluation theories in that it follows a factual perspective rather than an ethical or economical point of view. Sen constructs the idea of ‘Capabilities’ and describes the approach as an evaluative framework that assesses “the realised functionings (what a person is actually able to do) or on the Capability Set of alternatives she has (her real opportunities)*”* (Sen 2000, p75). Through the selection of factors grouped together within Sen’s conceptual construct called the Capability Set, one is able to enumerate the most valued aspirations of an individual and thereafter identify relevant indicators which are used to quantify the effective freedom a person has. Thus one is able to assess the degree of actual freedom the person has attained based on the selection of underlying indicators. In Ingrid Robeyns’ (2011) review of the Capability Approach, she described the approach as one that:

> Prioritizes certain of peoples' beings and doings and their opportunities to realize those beings and doings (such as their genuine opportunities to be educated, their ability to

> move around or to enjoy supportive social relationships). This stands in contrast to other accounts of well-being, which focus exclusively on subjective categories (such as happiness) or on the material means to well-being (such as resources like income or wealth (Robeyns 2011 p2).

In Sen's (2012, p47) 'Development as a Capability Expansion,' he contrasts the strength of the Capability Approach to general welfare economics that often identifies value in subjective measures based on the mental state of those that face deprivation, such as happiness or pleasure. Sen finds that such measures are misleading and do not reflect the real state of deprivation whereas the strength of the Capability Approach is in its analysis of what constitutes human freedom as this is the central feature of living.

The application of the Capability Approach is contextually grounded based on what a person can realistically achieve and aspires to achieve. The theory has an informational foundation, as the listing of freedoms can be empirically represented by available data which describes these freedoms. Through the definition of a Capability Set, one is able to identify the various functioning combinations that a person can attain. The choice of functioning represents a freedom that is attainable by a person. According to Sen, the simple task of keeping track of the most basic set of functionings such as being well nourished, escaping morbidity or premature mortality can have much greater value than the data reported on using a commodity based approach to welfare economics. As Sen (2012, p48) states, "The capability approach can, thus, be used at various levels of sophistication, and how far we can go would depend much on the practical consideration of what data we can get and what we cannot".

Four key concepts have emerged within Sen's work, viz., Freedoms, Agency, a Capability Set and Functionings. These concepts allow one to describe the underlying elements and relationships between these elements which constitute the Capability Approach. It is crucial, when unpacking these concepts, to view them from the affected individual's perspective as Sen's approach is targeted at individuals instead of a community.

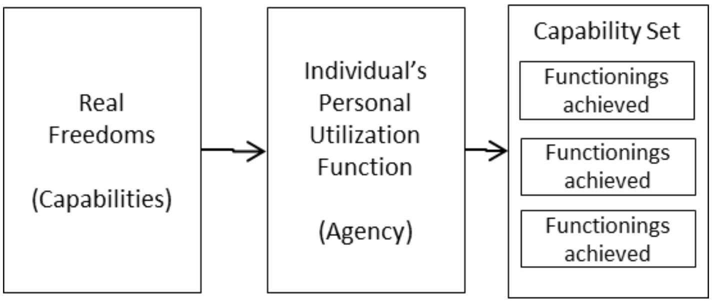


Figure 2: Relationship between Components of the Capability Approach

Freedom according to Sen is the actual real opportunity that an individual has to accomplish a task that they truly value. The choices one has, helps identify if a person is actually able to attain the ‘good life’ that one values, in contrast to the choices one is forced to make, due to the lack of genuine choice available to them. It is the assessment of ‘real opportunity’ that is critical in determining whether valued functionings, from the perspective of the individual under examination, can be achieved. Alkire (2002) emphasises that freedom must be understood in terms of the real-life possibility of attaining a freedom, as suggested by a nation’s legislation. Ingrid Robeyn’s (2011) describes an individual’s Capabilities as their real freedoms. In this regard there is a clear distinction between capability and functioning, as capability refers to the actual opportunities experienced in real life, whilst a functioning represents choices or aspirations that individuals value although it may not be specifically attainable. For purposes of this study, the identification of the central real freedom which explains why learners drop out of school is crucial, to begin determining which aspirations or choices would describe the multidimensional set of reasons for dropping out.

The concept of agency is best described in terms of how one is able to practically achieve a particular real freedom. This refers to one’s ability to act in pursuit of achieving a particular state of value (Alkire 2002). Agency is defined by Sen (1999 pp18, 19) as the action a person is free to accomplish in pursuit of their specific and personal goals. Furthermore, the agent is someone who acts in a manner that leads to change. These actions can be judged in terms of what the individual identifies as valuable. Therefore, Agency is the individual’s ability to act and achieve an objective that they value. In Robeyn’s (2011) review of Sen’s work she notes that the ideal of agency can be used to weigh the importance of a selection of applicable capabilities as the greater a person’s ability to act to achieve a particular state, the more attainable such a state is. In Flavio Comim’s (2001) explanation of the concept he notes that the agency aspect simply emphasises what opportunities or freedoms are available to an individual and in no way influences the extent of what that freedom is. For example, if a learner, the ability to walk to school, it does not guarantee that the learner will walk to and attend school. There are additional factors, one must consider why the learner does or does not attend school. In addition, the action that a person can undertake does not impact the range of desired states (functioning), one can achieve. Using the example of the learner again, the act of walking to school does not immediately ensure that the learner is able to attain all the aspirations one has from attending school. The limited effect that this concept has on either the Capability or Functioning has lead Comim to note that Agency is not crucial when operationalising the

Capability Approach. Following Comim's discussion, whilst the learner's act of achieving a particular state is important, it is not included as part of the assessment framework in this study.

The Capability Set is described as the various combinations of 'functionings' which are identified as valuable 'beings' and 'doings' that a person aspires to achieve (Sen 1992 p39). Sen referred to these combinations as a Capability Set and this array of functionings reflects an individual's freedom to choose what type of life they wished to live based on their own particular value system. It is the identification of these choices available to individuals, that is essential when constructing and populating the Capability Set. Functionings are the options or choices that an individual would value and therefore choose to attain. They are intuitive and have intrinsic value to an individual (Alkire 2002). Sen referred to Capability as "primarily a reflection of the freedom to achieve valuable functionings" and noted that functionings could be construed as the achievements of well-being (Sen 1992 p49). Functionings can further be grouped into two particular categories based on the characteristics they present. These include the choice of being in a particular state and the ability to perform valuable doings (Alkire 2002). As Robeyn's (2011) notes, examples of 'beings' include the valued choice of being well-nourished, healthy, housed, educated, literate, etc. and examples about the state of 'doing' include options to travel, voting in an election, attending school and moving freely, etc.

In order to apply the Capability Approach, it is vital to understand the central concepts of real freedoms versus valued functionings. By identifying indicators that quantify what is truly achievable (Real Freedom) in contrast to what the individual aspires to attain (Capability Set) illustrates the mismatch between reality and one's ideal aspiration.

3.1.2. The Progression of the Capability Approach

The Capability Approach has evolved since its introduction in Sen's (1985) 'Commodities and Capabilities'. It has been adapted and put into practice by a series of authors over the years. Since its inception, the approach has gained prominence due to the author's alternative views of standard economic frameworks and in the manner it places value on issues of poverty and inequality.

David Clark (2005) noted that Sen's Capability Approach has its roots in Aristotle's philosophies, Marxian Economics, Classical Political Economy and Rawl's (1971) Theory of Justice which refer to issues of dignity, self-respect, liberty and freedom. Sen's (1983, 1985) early work critiques traditional welfare economics as the field tended not to distinguish between well-being and opulence. Sen conceptualised the linkages between Commodity, Capability, Functioning and Utility, where commodities such as income enable capability (the ability to function) and recognising that capability enables the specific functionings that a person can access. The various functionings available to an individual provides them utility such as happiness. Sen notes that an individual's circumstances such as physical disability or other impairments limit an individual's capacity to transform commodities to capabilities. Therefore, the amount of effort required to achieve the utility such as happiness can be more complex and harder to reach. Various examples of utility exist and were highlighted by Sen in the early 1980s. He concluded this analysis by noting that neither the various commodities such as income or opulence nor utilities such as happiness, desire or fulfilment suitably represent the needs of human development, and rather a greater emphasis is needed on functionings and capabilities.

In Sen's (1992) next book, 'Inequality Re-Examined' in which he continues to advocate for equality and libertarianism. Sen further details his academic interpretation of the Capability Approach with a focus on capabilities and functionings and within this work he expands on the relationships between freedom, the concept of agency and well-being. Sen contrasts the traditional view of well-being to agency where agency relates to the manner goals and values can be realised and pursued and states that the concept of well-being should not be measured in materialistic terms but rather more in terms of how capable a person is to action personal objectives. Thus, agency is directly linked to well-being (Sen 1992 p56). Using a simplistic example, a blind but rich learner, may value sight more than money and may not personally believe they have achieved an advanced state of well-being. Although the text is deeply academic and philosophical, Sen's argument concentrates on the need to promote basic values such as individual capabilities and the freedom to achieve personal objectives which relates to the idea of Agency. John Broome (1993) in his review of the 'Inequality Re-examined' noted that Sen's argument to equalise capabilities ahead of income was very powerful as the ratio between income and the cost to meet an individual's needs differs across populations. Broome argues that through the better understanding of the ratio of needs contrasted against income would allow one to understand the importance of Sen's argument for freedoms and agency.

Sen together with Martha Nussbaum (1993) jointly edited their next release titled 'Quality of Life'. Working together with Nussbaum, following a conference on the international perspectives regarding the relative quality of life, the pair advocated for the systematic exploration of the wide range of dimensions related to the assessment of well-being, especially gender related concerns such as woman's exclusion from accessing opportunities and functionings. Robert Sugden (1994) in his review of the work noted that the indicators expressed differed based on the country or subject area that was under review. These distinctions highlighted how the relative sets of needs changed in relation to the environment. Underlying each exploration in the text was the unifying fact that the moral argument of Sen to consider capabilities ahead of factors such as income was necessary. The conference and the contributions within the text were some of the early practical explorations of the application of the Capability Approach.

Sen (1999 pp10-11) in 'Development as Freedom', identified social opportunities in the form of access to education and health facilities as one of the five types of freedom which he believed were instrumental to advancing the general capability of a person. These freedoms also included political freedoms, economic facilities, transparency guarantees and protective security. Sen believed if research on development shifted from a focus on income poverty to a more inclusive view of capability deprivation, one could better understand poverty from multiple perspectives (Sen 1999 p20). This text highlights the expansion of the Capability Approach from the perspective of freedom and libertarianism to have a greater inclusion of social elements, with an emphasis on its application. The various forms of freedoms are interconnected and when collectively advanced, greater resources for social opportunities are developed. Sen argues that although human rights are protected in countries' constitutions or enshrined in the United Nations Declaration of Human Rights, it is often impossible for such rights to be attained by all communities across nations (Sen 1999 p231). Paul Streeten (2000) agrees with Sen in his review of Sen's work when he states that often population growth is used to explain the harmful effects of poverty and that simplistic views of productivity and economic growth are extremely harmful when such approaches are adopted as solutions to problems such as poverty and inequality. Hence the need to identify and quantify the wide array of social factors which influence the capabilities of an individual, is critical. Sen argues that traditional economics negate the social concerns and therefore exclude such concerns from analyses of productivity and economic output.

Sen (2009 pp xi, 10, 96) released 'Idea of Justice' which returns to an analysis of John Rawls' discussion on justice. In analysing the various facets of justice, fairness and in determining how to assess such concepts, Sen introduces the use of a comparative perspective to assess the progress that different regions experience in fights against oppression or systematic social neglect. Mutee ul Rehman (2009) contends that justice is comparative in its very nature and supports Sen's assessment regarding the need for comparison to aid the identification of social ills and challenges, which is consistent with the need to identify comparable data amongst the BRICS.

Since the inception and revision of the Capability Approach, Sen has moved towards operationalising the theory. Although Sen contends that the factors that affect different regions and even individuals differ, the core concern that emerges and requires promotion and greater examination is one's capability. I.e. what determines one's level of real freedom. In order to promote capabilities; one must be able to identify suitable measures that represent the status of such concerns. Over the years, Sen has developed and clarified the concerns of capabilities, functionings, utilities, freedoms, well-being and agency. These concepts lie at the heart of describing the approach, however, once the approach is understood, it must then be applied. Together with Martha Nussbaum, Mahbub ul Haq and various other eminent authors, Sen has discussed how the Capability Approach can be applied.

Amongst the texts referred to in the above section, we note the movement towards identifying dimensions of well-being that refer to the various social opportunities that emerge. Sen broadly identifies access to education and health facilities as a central freedom, but does not delve into greater depth, especially in the education sector. These opportunities are closely aligned to freedoms and justice. In order to assess how well such justice or freedom is achieved, an associated appropriate set of data is identified to describe the provision of such services.

By reviewing the application of the Capability Approach in the next section, we can better understand what factors need to be identified to assess capability and justice.

3.2. Application of the Capability Approach

In Sen's (1985) most technical elaboration of the Capability Approach, 'Commodities and Capabilities', he highlights that underlying these concepts are the five key components that underpin the moral value of the theory. Firstly, it is important to accept that only real freedoms experienced by an individual produce advantages for them. Secondly, at an individual level there are differences amidst the population that affect their ability to transform resources into an activity that they value. Thirdly, the happiness of an individual is attributed to multiple factors that require individual recognition. Fourthly, in order to evaluate human welfare, we require a balance of both the materialistic and non-materialistic factors. Lastly, we must be aware that opportunities within society are distributed and not always equally accessible. Considering Sens's principles, any application of the theory must be conscious of these requirements which emphasise the importance of the impact on the individual, the value of materialistic and non-materialistic factors as well as the disparity in access of opportunities.

To explain Sen’s principles using a simplistic set of examples, in the order presented by Sen, (1) real freedom to access school is not possible, if a learner cannot afford the school fees to enrol at the school, (2) a blind learner does not receive the same benefit from a textbook as opposed to a learner who is able to see, (3) the learner’s happiness at school is a complex subject which is not only measured in getting access to the school, (4) school going expenses may not be the only factors which impede the learner’s appreciation of school and (5) a learner may have access to school in an urban area, whilst fairly limited access in some rural areas. Thus the provision of access to schooling and the benefits and appreciation derived from school may be experienced very differently amongst the population.

The Capability Approach’s strength lies in the manner that it supports a selection of indicators and how it can be put into practice or operationalised. Various attempts at operationalising the Capability Approach have been made by authors. The notable works are by Mahbub ul Haq together with the UNDP and Sen (United Nations Development Programme 1990), Sabina Alkire (2007), Alkire et al. (2008) and Robeyns (2011). Most quantitative applications of its use lead to the construction of a new survey, based on a set of functionings which describe what is most valued by the subject of the study. Note that the application of the approach in this study requires an examination of public secondary data and not the creation of a new survey. An important statement made by Sabina Alkire (2007) noted that the concepts of functionings and freedoms identified by Sen provide a basis for the comparison of well-being required for social evaluation. Therefore, one is able to begin to compare a complex phenomenon by itemising the core functionings and/or freedoms and related indicators which best describe those functionings or freedoms.

The most notable application of Sen’s Capability Approach was in the construction of the Human Development Index (HDI) as developed by the UNDP (1990) and the primary authors of the index were economist Mahbub ul Haq working alongside Amartya Sen. The index was ground breaking as it shifted the emphasis of countries’ state of development away from Gross Domestic Product (GDP) based measurements and towards a broader human well-being and standard of living base of measurement. Three important development concerns are included within the HDI in the form of access to health, education and goods. The HDI itself is a combination of indices such as life expectancy, literacy, school enrolment and income (Stanton 2007). Thomas Wells’ (n.d. p16) in his review of the Capability Approach and the

HDI states that the HDI does not fully reflect the scope or methodology of the Capability Approach but the HDI does succeed in introducing an alternative measure used in the assessment of economic metrics. With Wells' caution in mind, when examining the education related measures, one must also ask whether the measures for literacy and school enrolment are sufficient to describe the state of education in terms of the Capability Approach, considering the range of factors discussed in Chapter 2.

Alkire et al. (2008), when commenting on the HDI, notes that the Capability Approach is a valuable framework to evaluate well-being of a state and one of its strong qualities is that the application of the theory is useful when measuring complex and subjective concepts such as freedom, such as conducted in this study.

3.2.1. Mahbub Ul Haq's Human Development Approach

Mahbub Ul Haq's Human Development Approach is deeply anchored in Sen's Capability Approach. Ul Haq frames his approach around the opportunities available to people and the choices people are empowered to make (UNDP Human Development Reports n.d.). Freedom of choice underpins each of the concepts selected by Ul Haq and Sen (United Nations Development Programme 1990 p10) and went as far to define Human Development as a 'process of enlarging people's choices'. The Human Development Approach which led to the development of the Human Development Index (HDI), therefore focused on identifying and measuring the many dimensions of people's choices. In the UNDP report on Human Development, it was noted that human development is broader than the Capability Approach as the approach also considered how capabilities were used by people. The UNDP and Ul Haq believed the broad definition of Human Development allowed the HDI to better capture the complexities of human life whilst the counter argument that was made, was whether such a broad definition was conducive to measurement, quantification and comparison. Alkire et al. (2009 p34) however noted that it is easy to misconstrue capability and choice and states that the notion of choice excludes the issue of value. Alkire contends the value of the choice is very important when understanding capabilities as Sen described the selection of functionings as the valued choices one would make.

The Capability Set identified within the Capability Approach enables the approach to be adapted within the Human Development Approach. This required the identification of a series of indicators which best described the range of choices available to people, however when selecting these measures at an international level, it was found (in 1989) that the availability of comparable statistics precluded the comprehensiveness of the index (United Nations Development Programme 1990 p11). The Human Development Approach was as much a critique of traditional economic measures such as Gross National Product (GNP) as it was about measuring how people faired and their general well-being. When building the HDI, Ul Haq (1995 p46) reflected that the index had to overcome three difficulties that previously failed approaches did not address. This involved (a) not developing a composite index of the array of measures identified, (b) the methodological framework used to develop a composite index was not sound and (c) the composite measure that was produced needed to be rigorous enough to be considered an alternative to GNP.

In developing the HDI, Ul Haq (1995 pp 47,48) and the UNDP followed six principles when producing the composite index. Firstly, the index had to cover the concept of enlarging people's choices such as longevity, comfortable living, employment, quality of the environment and their freedom. Not each issue could be measured but the analysis requires one to cover these elements. The second principle was to ensure the methodology in compiling the index was simple and manageable. Therefore, the selection of the measure per concept was important and correlations between variables needed to be investigated and variables had to be discarded if their relevance was not clearly identifiable. Thirdly, a single composite index would be developed that would produce a single comparable indicator that could be considered an alternative to the GNP. Fourthly, the HDI would need to be inclusive of both social and economic concerns. Fifthly, the methodology needed to be flexible to allow for future adaptations and revisions. Lastly the sixth principle was to identify the most reliable and up to date data in order for the index to have the most value and impact as a policy persuasion tool. These six steps highlight the rigor that must be applied when applying the Capability Approach, the need to critically evaluate the usefulness of a particular dimension and the importance of collecting data of a suitable quality.

In Sen's (2000) review of the Human Development Report and Index, he comments on the pluralist approach to evaluating human progress. This idea correlates with the complexity of the lived experience that cannot be simply reduced to a single traditional measure such as GNP. Sen believes this plurality of the HDI approach is its greatest strength, therefore he contends that the indicators incorporated within the HDI need to present the multi-dimensional nature of the world, and should not present a view of being the single and only indicator taken into account in developing the composite index. The HDI highlighted the challenges of a complex world where there are multiple concerns that national governments must address. The nature of the HDI was to remain broad, inclusive and open to adaptation. Whilst, Sen makes these remarks in relation to the world's reliance on GDP as the only indicator on Human Development, one can draw parallels between this reliance and the countries of Brazil, India and South African which depend on the singular school dropout rate which do not reflect the pluralist nature of school dropouts.

The simplicity or naivety of the HDI formulae has drawn various criticisms. In response to these criticisms, Sen is quoted in his reply during an interview, "I want a measure that is just as vulgar as GNP except it is better" (Wallace 2004).

3.2.2. Application of the Capability Approach by the Human Development and Capability Association

The Human Development and Capability Association (HDCA) was launched in 2004 with Amartya Sen acting as the first president. The HDCA focused on multi-disciplinary research informed by the theories of human development and the Capability Approach with the intention of bringing together the numerous academic networks working in this space (Human Development & Capability Association n.d.).

Sabina Alkire and other members of the HDCA (2009 p40), have noted of the value of the Capability Approach in its application and resultant empirical analysis that it enables, however the key questions that Alkire as raised are (1) how does one identify the selection of capabilities that are valued by people and (2) when evaluating a particular policy, which capabilities are actually relevant? These questions correspond with the six principles discussed by Ul Haq and draw attention to a gap that exists between national policy priorities and an individual's freedoms that are experienced in reality. Whilst this can be considered to be a challenge in policy analysis, the contrast of the real freedoms versus policy priorities could raise pertinent concerns for governments to tackle.

Alkire et al. (2009 p106) shifted from a broad inspection of human capability and focused on poverty and inequality alone. The authors tested the practicality of the theory with respect to various indicators of poverty and inequality and highlighted the strengths and weakness of these measures but noted much of the research pertaining to indicators often identifies functionings without properly measuring freedom or capability. The methods that the authors proposed included the following, (1) if the selected measures represent a selection of functionings, the collective aggregations would represent the impact on the freedom experienced, (2) if the intention is to measure whether ones actions are coerced or were taken freely, it would be useful to map the indicator of coercion against the particular functioning that is under review, (3) if the research has an economic perspective, it may be possible to map a capability against an equivalent income for a particular region, (4) data regarding perceptions of one's opportunities in a particular area could be used as a direct proxy of one's capability and lastly (5) in the absence of capability data relevant to the functioning under review, data on income and time use in relation to the functioning could be adopted. Alkire' et al.'s five methods for applying the Capability Approach highlight lessons that have been learnt when linking data to identified functionings. Furthermore, this highlights that the selection of functionings and capabilities are the critical step to enumerate the factors that best describe a phenomenon such as school dropouts. In addition, points 1 and 4 are highly pertinent in this study, especially in consideration to the structure of the identified Capability Set as discussed later in section 3.4.3.

Alkire et al. (2008 p5) further expanded on the data challenges angle and noted that when most surveys or data collection instruments are formed, the aim of their collection was not to collect information on functionings, well-being or capabilities. Therefore, care must be taken to apply the correct selection of variables against functionings or capabilities when attempting to perform an empirical analysis informed by Capability Theory. Whilst understanding there are data limitations, Alkire et al. still further highlighted the uses of the Capability Approach to date and identified its use in (1) assessments of countries' human development state, (2) assessing smaller scaled development projects usually as an alternative to cost-benefit analyses, (3) identifying the location of the poor in emerging economies, (4) profiling the location and challenges of the poor in advanced countries, (5) quantifying the deprivation of disabled people especially considering that standard income analysis does not represent the disadvantages that the disabled experience, (6) in the assessment of gender inequality, (7) providing a means to surface inequalities and thereafter debate the necessary policies required and (8) providing a means to critique existing norms and policies. Each of these uses is relevant when critiquing the school dropout phenomenon in three geographically dispersed regions as this study contends. Varied examples of such implementations and the identified set of relevant capabilities and functionings are discussed in the following section.

3.3. Key themes of Human Development using the Capability Approach

Amongst the applications of the Capability Approach that were discussed above, the studies with a direct bearing on this study is the work of Unterhalter (2009), Unterhalter et al.

(2015), Walker (2015) and other notable authors that have applied the theory to the education sector. These studies assist in identifying which themes are most valued by learners. However, considering the broad range of factors reviewed in the school dropout literature discussed in Chapter 2, one notes that the learner's concerns are complex. They are not only determined by factors and experiences from within the school, but also those from the household, the community environment and other general human welfare perspectives. Considering the relevance of these broader perspectives, the themes that have emerged from human development applications, highlighted by Capability Approach experts such as Sen, Ul-Haq, Nussbaum, Alkire, Cummims, Narayan and Comim in will help shape the structure of the School Dropout Capability Assessment Framework which follows in Chapter 6. The following section unpacks the education specific themes and thereafter the broader Human Development themes which are relevant for this study.

3.3.1. Education sector specific themes

Elaine Unterhalter is an internationally prominent Education/Capability Approach researcher whose analysis of Sen's exploration of the Capability Approach as applied to the education sector found that the Sen's deliberations on Education seemed contradictory to his actual views in applying the approach. Unterhalter, in her Chapter in The Capability Approach: Concepts, Measures and Applications (Comim et al. 2008 p490), asserted that Sen often presented education in very homogenous terms instead of a multi-dimensional, heterogeneous approach for which Sen was known to advocate. Unterhalter believes that assuming there is uniformity in the education system is problematic as there is great empirical evidence that there are numerous capability deprivations that affect schooling. Unterhalter also notes that in some schooling systems in the world, the education system actually contributes to repression rather than enabling freedom. In this light, there are many contextual considerations regarding how education should be analysed. Unterhalter also notes that Sen's assertion that the simple provision of basic education to act as an enabler for capabilities is not sufficient and that greater emphasis is needed on quality education that encourages learning rather than a casual linkage between education and capabilities. This finding supports the need for a deeper analysis of issues which affect the learner.

Melanie Walker (2005b) found the Capability Approach useful in education analysis and noted that "Sen's Capability Approach offers an approach to evaluating social (and, hence, educational) advantage, in which expanding people's agency and freedom is held to be central"

(Walker 2005b p104). Walker together with Unterhalter and Vaughan (Unterhalter et al. 2015) suggested that the analysis of the education space using the Capability Approach was still a developing area of theory that was subject to various debates regarding its interpretation. Often these debates were due to differences in opinion in applying the theory by following either Amartya Sen's viewpoints or Martha Nussbaum. This debate followed from the contestation of viewpoints between a prescriptive list of broad freedoms necessary to describe the education sector (Nussbaum's perspective) or a nuanced analysis of freedoms and functionings relevant to an individual's personal challenges and aspirations (Sen's perspective). The approach followed in this study, has been to explore the many angles discussed in Capability Approach and school dropout literature, applicable to the three countries of this study and those raised internationally, to produce a finite list (following Nussbaum's method) that is nuanced by the broad set of concerns discussed by school dropout experts (incorporating Sen's argument for nuance when applying the theory).

Recognising the limited scope of capability analysis in the education sector, Unterhalter in the Alkire et al. book (Alkire et al. 2009 p207 - 227) delved into the theoretical complexities and argued that education is a core enabler of real freedoms and plays an instrumental role in uplifting economic growth, therefore in the education field specifically, it is critical to understand the capabilities and functionings enabled by the provision of the service that is studied. Unterhalter argues that education fosters knowledge propagation, enables empowerment and is central to human flourishing by opening gateways to other valued capabilities. She continues by citing Sen's discussion of how education supports the expansion of capabilities due to the instrumental social role it plays in communities in general, in improving community participation in governance and the empowering and distributive nature it has in facilitating marginalised communities to get access to centres of power. Therefore, when one is identifying the real freedoms, one must consider the gateway to other valued capabilities that education offers such as employment opportunities and a better quality of life.

Unterhalter's categorisation of the types of applications of the Capability Approach in the education sector highlighted three groupings, (1) studies that highlight the value of education and provide options for its delivery to communities using the terminology of the Capability Approach, (2) the various studies that considered the intersections between human rights, equality, social justice and education and (3) analyses of peoples' view of learning and the value it presents in their lives and the measurement thereof, of these views (Alkire et al. 2009 p207 - 227). In studies that used the Capability Approach terminology, it was often asked whether the approach could also be used as a means to identify where there was a linkage between input resources into the school and the capabilities that they produced. Such analysis would be useful in terms of a policy review. In consideration of the above, one may note that there is a gap in education analyses which targets a particular phenomenon of school attendance (and the consequent dropout) as conducted in this study, thus also highlighting the need for a deeper form of analysis.

Unterhalter et al. (2015) notes that Education Capability Approach research to date has been exploratory in attempting to identify valued functionings and capabilities appropriate for the sector. This exploration of the sector proposed the following vital factors, viz., class participation, learning about history, a learner's ability to interact in class, developing vocational skills, numeracy, general knowledge and confidence. However, authors are undecided in determining whether the research is comprehensive of all education concerns and whether such an approach of itemising such factors follows too closely from the Nussbaum theory which prescribes a list of central functional cored education capabilities. Ultimately Unterhalter contends that the primary reason that the Capability Approach is appropriate for the needs of the education sector is that it helps one identify the sets of measures that "question the range of real educational choices that have been available to people; whether they had the genuine capability to achieve a valued educational functioning" (Unterhalter et al. 2015 p2). These questions are useful to help establish goals related to school attendance and reasons for dropout across the BRICS countries and specifically for Brazil, India and South Africa. In addition, there is a need to recognise the differences in needs of individuals, where for example a disabled learner has a different set of needs to an able-bodied learner.

Madoka Saito (2003, pp24, 25) in his exploration of the Capability Approach also underpinned the approach's usefulness to the education sector noting how the HDI was instrumental in building awareness of the importance of education across countries. Saito further made the connection between human capital and human capability and found that human capital inputs into the labour markets received from education could also assist in uplifting human capability. For example, learners who attain a higher level of communication skills from school participation can be found to be more productive in employment and prosper in general when compared to those without such skills. Saito asks why learners value attending and completing school and believes it is due to their need for suitable qualifications which enables them to productively participate in the labour market. This corresponds with

Unterhalter's argument for recognising the gateway offered to learners by education to access other valued functionings.

Rene Vermeulen (2014) suggests that the value of the Capability Approach in Education is due to the recognition it offers to more than just the economic value of the sector in that one's development perspective should not only be limited to how productive an employee a learner will become when they enter the job market as suggested by Saito. Vermeulen was of the opinion that the approach reinforces the view that education should be considered a human right and therefore found great value in its incorporation within the Millennium Development Goals (MDGs) (Goal 2). However, a focus on indicators which only emphasise the right to access to education (as done within the MDGs) still leaves a gap in when assessing the quality of the education offered to learners. Vermuelen believes greater emphasis must be directed to student learning rather than simply the attendance of class. Furthermore, she argues that the common indicators of education quality such as pass rates and scores from standardised tests often lead to a simplistic view of education quality. Vermuelen refers to the UNESCO framework to better understand education quality as an option to better analyse the education processes in terms of the learner's input, the enabling environment and the outcomes from the process. Vermuelen believes it is the strength of the Capability Approach that will assist in analysing, monitoring and evaluating the quality of education in a manner that traditional economic analysis may find irrational (Vermuelen 2014). Vermuelen's considerations are important when one assesses the quality of education received which, in its own right, is also a complex multidimensional concern when providing education.

When applying the Capability Approach specifically to the school dropout phenomenon and school attendance in general, the push-out factors described by Cardoso and Verner (2006), and the risk factors described by Chugh (2011) will need to be unpacked. Questions of whether early parenthood, child labour and the effects of poverty or family background and domestic problems require an assessment to determine if they feature as of functionings or capabilities relevant to school attendance. It must be determined how such factors contribute to a learner's decision to abandon attending school.

Unterhalter outlines one of the central debates pertaining to the use of the Capability Approach within the education sector. Questions are asked whether a predefined broad list of capability related indicators best describe the education sector or if a nuanced approach dependent on the context of the situation lends itself better to this form of analysis. Unterhalter believes making such a choice is not necessary as both scenarios can be useful. Importantly we find that identifying the key range of education choices is what underpins the usefulness of the approach in the education sector. The idea of personal choice differs from traditional economics that associate education as an input to improving the level of human capital available to the labour market. In fact, the approach is far more useful and can be used to assess education quality and the wider education system as suggested by Vermeulen and carried out by Cardoso and Verner (within section 2.3) who have actually applied the theory at a more granular level in Brazil.

In review of these factors discussed in education related Capability Approach applications helps provide the outline of what are the real freedoms gained from school attendance. However, dropping out of school effectively curtails any benefits attributable to school attendance and completion. Thus the learner's effective freedoms, derived from education access, is linked to the reality of their circumstances which are broader than factors related to learning or the school environment. The broader factors are influenced by general human development concerns must be assessed to identify all the states of being and doing that learner's value.

3.3.2. Human development themes

The many experts involved in Capability Approach assessments of Human Development have presented various applications of the theory. They identify the broader elements of a framework following experiences they face within their homes and communities. Many authors have highlighted the key themes which have emerged which describe human development which is inclusive of financial security, well-being, safe travel to school and the ability to express oneself freely amongst various other themes.

From Sen's (1999 pp1-11) perspective he identifies the provision of education as a central enabler of capabilities of a person to achieve their valued aspirations in life, which corresponds with Unterhalter's position on education analyses. Furthermore, social opportunities such as education and health enable a person to be a productive participant in the economy at a later stage in their lives. Therefore, education is viewed as a gateway to achieving a valued outcome and is not an outcome on its own.

Mahbub Ul Haq (1995 p27) identified the effects of poverty as a central impediment to human development which needed to be addressed directly. A reduction of poverty enables citizens to better provide for themselves individually strengthening their ability to access better quality services. This factor is especially valid in terms of the costs of education and its related set of supporting requirements. Ul Haq argues that poverty alleviation is central to human development and this includes the need for access to income, assets, credit, social services and employment. Ul Haq also found that the provision of an enabling environment is essential to human development. An enabling environment typically promotes people's access to infrastructure and resources, allowing individuals to perform their valued tasks. Furthermore, the environment does not constrain people's expression of political or cultural convictions. According to Ul Haq a negative environment has a limiting effect on people's choices and reduces one's access to knowledge, nutrition and mental health.

Martha Nussbaum (2000 p78 - 81) believed it is important and necessary to find a set of central capabilities that affected all people. Her list stemmed from humanities common appreciation of what it meant to be truly human. Nussbaum proposed a list of ten central human capabilities which she argued were vital, considering that all people are first human and therefore share an equal appreciation for these freedoms as humans. This list represents the tenets of good living that all people valued. This included being able to live a life of normal length, being able to have good health, to be able to move freely without risk of violent assault, being able to imagine, think and reason, being able to love, grieve and freely experience the full range of human emotions, being able to engage in critical reflections openly, being able to live with and towards others by being able to recognise and show concern for others in social interaction, being able to live with other species and nature in general, being able to enjoy recreational activities and to be able to participate freely in political choice that govern one's life. Although Sen supports the pluralist view of applying the Capability Approach and avoided prescriptive tendencies when applying the approach, he supported her setting of a fixed list noting that it was important for studies to identify and share best practices when identifying important capabilities (Sen 2004). What Nussbaum was able to identify a common set of capabilities that were applicable in all settings. Where any of the capabilities are curtailed, a person's ability to live a truly human life would be diminished. Therefore, Nussbaum's identification of ten capabilities is the foundation of most Capability Approach assessments. Of specific importance in this study is Nussbaum's identification of bodily integrity as a central human functional capability, which refers to one's freedom to move from place to place and in the context of this study, the freedom is relevant to a learner's commute from home to school and back. Such freedom relates to one being free from harm, be it assault or abuse.

Robert Cummins' (1996) argued for the use of 7 domains of life satisfaction after exploring 27 definitions of life quality. Through the application of the Capability Approach, Cummins identified material well-being amongst other factors such as being healthy, being productive, intimacy, safety, community and the feeling of emotional well-being. Cummins notes that material well-being is a necessity for a comfortable life within 59% of the studies that he reviewed. It has a direct linkage to one's socio-economic status and is recognised as an enabler to access varied commodities and services. By understanding that families require the financial security to access and/or possess items such as a home, clothes, food and to attain a certain standard of living. These are essential to acquire the necessary resources for the household which enable learner's and their households to attain other valued states of being or doing. Although various capability researchers avoid including financial security as valued functioning as it is often seen as a gateway to other valued states, and having material wealth, is not seen as an end result. It is Cummin's analysis of human welfare and survey of such literature which has led to his inclusion of this domain.

Whilst Cummin's surveyed human development literature, Narayan et al. (2000), from the World Bank, conducted an extensive study of poor communities across 23 countries to identify their most valued states of attainment. Similar to Cummins, Narayan et al. also emphasises the importance of physical and emotional well-being, amongst other distinctions in well-being. They argue that poor communities differentiate well-being from wealth by highlighting that financial security neither guarantees one security nor respect in their communities. In expanding on what well-being refers to, Narayan et al. describes physical well-being as inclusive of bodily well-being, which refers to one's health and access to health services. The authors find this dimension as essential if one attempts to explore the needs of a good life which all people aspire to have. For example, although material well-being is essential, one may find it difficult to access to employment if their bodies are not healthy or strong. Therefore, bodily well-being is an enabler to material well-being. Narayan et al. also contends that one's psychological well-being, expressed in terms of one's peace of mind, lack of anxiety, happiness and personal harmony, is essential to material, bodily and social well-being. Such peace of mind is essential to attain the focus needed to complete the valued task hand, however, at the same time, feelings of humiliation, shame, anguish and grief jeopardise the person's state of mind and weakens their ability to perform effectively in their chosen role. The effects of poverty contribute to such distress.

Alkire et al. (2009) discuss the use of infrastructure as an enabler of a person's basic needs. These basic needs represent the goals for development which allows individuals to live a full and decent life. The authors also note that access to infrastructural assets is a key driver of human development and within the education sector, access to schools and the resources within the school enables a person's attainment of necessary skills and knowledge to perform productively in society. Limitations on one's access to infrastructure, therefore limits a person's ability to realise the valued goal of education attainment as the learner is unable to convert the use of such resources into capabilities and/or functionings.

Santosh Mehrotra within the Comim et al. (2008 p386 - 417) study emphasised access to basic services as a key driver of the Capability Approach and thereafter discussed how the limited provision of schools within communities disempowers the learner due to their limited ability to participate in schools. Furthermore, Mehrotra notes that if a learner was never able to attend school in the first place, the learner is not considered as a dropout and is simply not factored into the scope of school attendance problems. Therefore, the location of the school is a critical enabler of school enrolment and participation.

The themes that have emerged from the Capability Approach literature is instrumental in establishing the framework required to assess these factor's effect on learners. Each of the broad themes represent states that are valued by learners but require greater definition in terms of their elements which were discussed in the school dropout literature within Chapter 2. The combination of school dropout literature together with the identified capability concerns will provide the necessary detail to describe the factors of school dropout. The next section discusses how one can transform these identified traits within the theory into the foundation of the School Dropout Capability Assessment Framework.

3.4. Translating the Capability Approach Theory into Practice

When conducting a Capability Approach assessment of a particular phenomenon, two crucial tasks emerge from the literature. The first step requires one to specify the key concepts that describe the phenomenon as identified within Capability Approach Theory, as discussed in the preceding sections of this chapter. These concepts drive the identification of core descriptive themes for each identified capability and/functioning. The next step involves moving away from conceptualisation of philosophical themes and associate practical existing datasets to each identified descriptive theme as noted by Comim et al. (2008 p158). Comim et al. describes the traits of the descriptive themes and then shifts towards identifying the practical concerns when implementing a Capability Approach application. Comim et al. thereafter identifies the steps that are typically following when implementing such an assessment. These steps are discussed in greater depth in Chapter 4 in the study's methodology.

3.4.1. Linking the Practical to the Conceptual

Comim et al. (2008 p166-185) shared Ingrid Robeyn's belief concerning the difficulty and challenge in operationalising the Capability Approach which they found lies in its measurement. A narrow measurement of a broad phenomenon seems counterintuitive considering Sen's emphasis on plurality as the individual has a desire for an array of outcomes. To ensure that the study has ensured coverage of the broad conceptual needs, Comim et al. has outlined a set of criteria that can be used to assess the implementation. This approach is incorporated in the testing of this study, which is provided in Chapter 8.

Conceptual Concerns	Practical Level
Valuational Foundation	Data
Human Diversity	Multidimensionality and Aggregation
Objectivity	Weighting and Incompleteness
Counterfactual Nature	

Table 9: Understanding Comim's conceptual and practical concerns

When clarifying the concepts to be used, Comim et al. (2008 p162 - 166) states that a 'valuational exercise' that must be conducted, noting that the measurement of capabilities should be informed by the principles of the Capability Approach. This covers a range of issues related to social concerns such as poverty, inequality, quality of life, social justices and social arrangements. Each of these concepts needs to be evaluated to determine which of these broad concerns are relevant for the study. If the assessment is meant to influence a policy, Comim et al. argues that the approach must also cover concepts such as opulence, utilities, primary good, rights, functionings and capabilities. Thereafter the researcher must itemize and weight valuable things that the people are able to be and do. The critical concept in this exercise is to identify what actually is valuable amongst these highly heterogeneous variables. Trivial concepts are identified and discarded. Sen (1992 p44-46) notes that the selection and ranking of variables is critical. Such analysis and decisions require public participation to assist in making such choices. In making these choices, in this study, the selection of capabilities and functionings will be informed by the key issues discussed within school dropout literature.

Comim et al. (2008 p166-167) contends that 'pervasive human diversity' is another major feature of the Capability Approach. These variations among people include personal heterogeneities such as general demographics, environmental diversities such as political persuasions, differences related to social culture and norms. These differences could be expressed in biases which may impede the capabilities of others. Each of these differences could materialise in the form of an individual's available resources which can be transformed into realised capabilities for the particular evaluation. Therefore, when measuring and evaluating the selected Capability Set, it will be important to list such factors and thereafter disaggregate these measures by the various diversity factors raised above.

The third key conceptual concern that Comim et al. (2008 p171-173) raises is in support of Sen's notion that assessments using the Capability Approach needs to be informed by objective measures rather than subjective measures of well-being or happiness. It has been found that restricting the application of a Capability Approach evaluation to only objective measures can be difficult especially considering how often analyses is offered using surveys that utilise subjective questioning to measure capabilities. In recognition of this problem Comim et al. states that when conducting a historical analysis of the progression of capabilities, one could study how such subjective measures fluctuate over time. Comim et al. notes that objective empirical work will be difficult when we consider the varying contextual spaces which affects the measurement of capabilities.

The last conceptual concern of Comim et al. (2008 p174-175) is the often counterfactual nature of empirical analysis. Often the distinction between capabilities and functionings is not clear in Capability Approach analysis and the analysis that emerges does not recognise when one's available capabilities are low but their actual resultant functionings are high. The example of Sen's that Comim et al. refers to, is one where an old person attains good health despite the various challenges that affected them in their lives. In this regard the outcome of health capabilities was counter-intuitive to the actual outcome/functionings of the older person. Such examples may be statistical outliers and it is still critical to identify the key enabling capability factors which influence a community. Sen and Comim et al. argue that in order to negate the counterfactual effects of empirical analysis, one must closely examine the link between a counterfactual question and the possible realisations. Although such questioning adds an extra layer of complexity to the analysis of capability, it highlights the need to critique the data that is selected and also ask whether individual capability choices impedes the degree of freedom one has made in lifestyle choices.

In terms of the practical requirements for this analysis, Comim et al. highlights three important concerns viz. (1) the sort of data available, (2) the manner the aggregation of multidimensional variables is applied and (3) how one weights these multiple variables whilst recognising that there is always space for additional or more accurate data.

In Comim et al.'s (2008 p177) reflection on the issues of available data, he notes that the majority of efforts when matching empirical data against an analysis of the Capability Approach usually focused on the wide range of relevant functionings with a limited inclusion of capabilities. This exclusion is often due to the inability to define capabilities. Three important sources for data according to Sen and supported by Comim et al. are financial data related to the market that is under study, surveys of public perceptions and non-financial data regarding an individual's personal status in response to the subject area. Often the non-financial data is best suited to Capability Approach analysis. In South Africa, as Klasen et al. (Klasen et al. 1998) notes, such data can be sourced from national statistical provider's range of household surveys which provide information on living standards measures. Sen (1992 p53) notes that international comparisons have been found to be difficult often due to the lack of comparable data. Sen contends that there is value in being cautious in which dataset is used. However this must be balanced against a need to accept the data that provides the best international estimate (Comim et al. 2008 p178). In Sen's 'Inequality Re-examined' he argues, that the Capability Approach can be used at varying degrees of sophistication in terms of the level of detailed data that are used in the analysis. The primary concern which guides the depth of analysis is the practical consideration of what data is available. This statement highlights the very pertinent question of how one critiques the quality of data that is available, which is consistent with the premise of this study involving the testing of data quality of public datasets. Furthermore, the level of detail needed in measuring capabilities would be guided by the level of analysis required. For example, a macro form of analysis would require very little information regarding interpersonal differences.

The second practical concern that Comim et al. (2008 p182) raises relates to the multidimensionality and aggregation of the selected data. Aggregations of data may seem contrary to Sen's (1992 p86) guidance on need for the data to be presented in heterogeneous manner and the very nature of the Capability Approach requires one to collect data across multiple heterogeneous variables. Therefore, the differences within the data need to be protected when the analysis is presented. Aggregating multiple indicators into single index could hide the complexity relevant to the analysis and therefore the analysis could discriminate against those performing at the lower range of the particular indicator. Sen is aware of this challenge and notes the level of aggregation employed should be dependent on the level of analysis that is required. The decision to collapse multiple indicators into a single index or aggregate variables across geographies or population demographics should be made in line with the purpose of the study. For this study, the selection of multiple indicators is not intended to be merged into a single index.

In terms of weighting the data selected, Comim et al. (2008 pp 184, 185) notes that the decisions made regarding the importance of the variables featured within the analysis are made in response to practical requirements. There are various techniques that can be adopted to help inform how a weighting decision is made. However, the employment of such a technique must follow participatory and democratic principles in how such reasoning is applied. The manner in which such weights are introduced, especially when applied to public policy, should be conducted openly and explicitly.

The degree to which these factors have been applied consistently in this study are described in Chapter 8. The following 2 sections of the chapter summarises the key statements within the Capability Approach literature, which form the outline of the School Dropout Capability Assessment Framework.

3.4.2. Defining Real Freedom: Learner's real freedom to complete school in preparation to access employment opportunities and an improved quality of life

Real freedom, as was discussed in section 3.1.1 by Sen, was described as the actual real opportunity that an individual has to accomplish a task that the person truly values. If we analyse this statement, there are few key terms that must be discussed and interpreted for purposes of this study, viz., actual real opportunity, individual and the tasks the person truly values. These terms must be interpreted in terms of the objectives of this study.

Sen's approach is applicable to all aspects of Human Development, whilst this study has a narrow perspective related to only to the determinants of school dropout within Brazil, India and South Africa. To do so, one must consider the impact that the lack of school attendance has on the learner's future opportunities. Following the arguments by Spaull (2015) and other authors, learners attend school to access a pathway to productive employment opportunities which enables their improved quality of life. Those that leave the school system fall off that pathway and find it more difficult to access suitable qualifications that are recognised within the labour market to access the productive employment opportunities. As Spaull states:

> Those 18-24-year-olds who do not acquire some form of post-secondary education are at a distinct economic disadvantage and not only struggle to find full-time employment, but

> also have one of the highest probabilities of being unemployed for sustained periods of time, if not permanently. (Spaull 2013 p3)

For the purposes of this study, the key aim is to identify which factors place the learner on this pathway to attain qualifications which enable employment and the desired quality of life.

In this context the actual real opportunity is framed within a narrow context of the learner attending school. Therefore, the act the learner truly values is attending school and therefore the assessment that is needed refers to what opportunity the learner truly has to attend school. This includes the choice the learner makes to leave school which is based on the demands of the learner's reality. As discussed by Cardoso & Verner (2006) and Reddy & Sinha (2010) within sections 2.1 and 2.2, school dropout factors are often referred to as push-out factors, which indicates that the reasons for leaving school are often beyond the learner's control.

Elaine Unterhalter et al. (2015) (as discussed in section 3.3.1) found that the primary purpose of applying the Capability Approach to the education sector was to identify the "range of real educational choices that have been available to people" (Unterhalter et al. 2005, p2). With a narrower scope focusing on learner's choice to attend school, there is only a single real freedom which can be assessed and this is whether the learner has an actual choice in attending school. Therefore, the study must assess whether the learner's reality empowers the learner to participate freely within school or whether their socio-economic conditions detract from productive participation. Despite prescriptive declarations in government policy regarding compulsory schooling, the learner's real choice in attending school can be found to be the net effect of a combination of socio-economic factors across the countries at individual, household, school or community levels.

In addition, this real freedom can be qualified in terms of the gateway functionings such as how a learner's completion of school enables her employment opportunities and an improved quality of life. Considering this clear linkage, it is important to connect the central freedom that learners aspire to have to the need for employment and an improved quality of life.

The identified real freedom, i.e. 'The Learner's real freedom to complete school in preparation to access employment opportunities and an improved quality of life,' as discussed by Sen in 3.1.1 of this chapter, can be best described by a collection of functionings which make up the Capability Set. This Capability Set includes the set of functionings which inform the learner's choice to leave school and can thereafter be qualified by the selection of indicators which describe the learner's valued state of achievement (being) and action (doing). The broad set of functionings appropriate for this study are discussed in the next section.

3.4.3. Identifying the School Dropout Capability Set

David Clark (2005) argued that an empirical approach to measuring concepts such as well-being and development is required. In Clark's application in South Africa, he found that that employment, housing, education and income amongst various other poverty assessments were found as the most pertinent aspects of a good life which was found to be consistent with findings by authors such as Sen and Nussbaum. Clark noted that an important clarification that was needed when applying the Capability Approach was detailing (1) the practical issues related to survival and development in emerging economies, (2) the mental functioning of human well-being and (3) an identification of the enjoyable facets of life. Furthermore, Clark notes, that when the approach is turned towards an analysis of public policy (as in this study), the expansion of the Capability Set is still required, however the focus is not on pure economic growth but rather to concentrate on widening the selection of basic capabilities such as employment, greater levels of prosperity and better provision of social services. Clark's review provides clarity in terms of how to collectively represent the plurality of a particular phenomenon that is tested in terms of generic umbrella items, which can be expanded in greater detail. Using the various studies referred to in the previous sections, one is able to identify the core set of functionings that make up the School Dropout Capability Set.

In review of the various themes that were discussed in section 3.3, the identification of functionings which make up the Capability Set must encapsulate the primary concerns highlighted within the literature regarding the school dropout concerns. However, these concerns must be assessed in the manner that Sen (1992), Nussbaum (2000), Alkire et al. (2009), Clark (2005) and others have adopted to enumerate the related functionings. Clark provides a rare analysis of the qualities that Capability Approach dimensions should possess when applied to a particular issue. These traits include the needs for survival as experienced amongst emerging economies, the mental state of those analysed and the experiences that are enjoyable facets of life. If we apply Clark's characteristics of dimensions to the school dropout phenomenon, we can extrapolate the following core dimensions which could form the basis of the Capability Set for this study. Each of these dimensions as expressed in Table 10, require clarification in terms of how they apply to learners dropping out of school and how they relate to dimensions used in previous by Capability Approach applications.

Clark's Characteristics	**Capability Set Functionings**	**Supporting Author**
Survivalist development requirements	Financial Means	Ul Haq (1995) – Reduction of Poverty; Alkire et al (2009) – Equity, Poverty; Cummins (1996) - Material well-being
	Physical Health	Nussbaum (2000) - Good Health; Narayan et al. (2000) – Bodily Well-being
	State of Infrastructure and supporting resources	Alkire et al. (2009) – Basic Needs Approach
	School Travel	Comim et al. (2008) – Access to basic services; Nussbaum (2000) - Move freely (free from violence)
Mental Well-Being	Mental Health	Narayan et al (2000) - Psychological Well-Being
	Learning Culture and Environment	Ul Haq (1990) - Quality of the Environment
Enjoyment/Appreciation	Effective School Participation	Ul Haq (2000) Education; Robeyns (2011) - Education; Alkire et al. (2009) – Participation and Empowerment
	Free Personal Expression	Nussbaum (2000) - Engage in Critical Reflections; Alkire et al. (2008) - Analyse in terms of demographics

Table 10: Core Dimensions of Capability Approach Assessment Framework

Through the synthesis of learner dropout factors as described in Table 10, we are able to determine the most valued set of influences on learners in the three countries that correlate to the set of core dimensions which make up the Capability Set for this study. These factors identified in the school dropout literature have been grouped together under broader umbrella items which relate a state of achievement (being) or a learner values or a state of action (doing) as Sen prescribes for the Capability Approach.

With respect to Clark's (2005) survivalist development requirements, the first functioning identified relates to the learner's feelings of financial security. This concern has been identified by various authors such as Ul-Haq (1995), Alkire et al. (2009) and Cummins (1996). Notably, as discussed in sections 3.2.1 and 3.3.2, financial means has been identified as an enabler of various other valued functionings, where for example, learners aspire to have greater resources for acquiring learning support materials or school uniforms, or being able to pay for their school fees. Thus, whilst some capability authors tend not to recognise financial well-being as a functioning, it is found to be highly pertinent when analysing determinants of school dropouts. Another survivalist trait relates to the physical well-being of the learner. Naryayan et al. (2000), as discussed in section 3.3.2, for example identifies bodily well-being as an enabler to financial well-being as an unfit person is unable to secure employment and similarly, an unfit learner is unable to productively participate in school. The third survivalist functioning is described best by Alkire et al. (2009) (also discussed in section 3.3.2) who recognises the state of infrastructure that an individual must contend with. Alkire et al. notes that the lack of access to basic infrastructure services limits one's abilities. When applied within the context of a school, a dilapidated school tends to leave the learner feeling unsafe or distracted from study due to the lack of basic services as well as a lack of access to school facilities and services. This includes the resources that support learning in the school such as libraries, laboratories, textbooks and other school related supplies. The last survivalist functioning involves the manner in which learners travel to school as discussed by researchers like Comim et al. (2000) and Nussbaum (2000) (see section 3.3.2). Nussbaum advocated that all people should feel free to travel and move without concern in public spaces. This she recognised as one of the traits which made a person feel truly human. However, the literature tends to indicate, that the learners that travel greater distances, not only suffer due to exhaustion but also face various perils on the journey to school causing the learners stress about getting to school.

In terms of mental well-being as identified by Clark (1995), the first factor which impacts the effective learning ability of the learner is their mental health. A key proponent of this factor is Narayan et al. (2015) who recognised the effects of the various forms of well-being including psychological well-being (see section 3.3.2). Considering the discussion in 2.4 by Thurlow, Sinclair and Johnson (2002), one also notes the deep concerns that relate to learning impediments and their limitations to learning. Furthermore, learning culture and environment, as identified by Ul-Haq (1995), (who discussed the quality of the environment that the individual must contend with) can be a limiting factor if there are influences within the environment which tend to pull the learner away from school access.

Clark's last characteristic is the enjoyment and/or appreciation a learner derives from school attendance. Robeyns (2011) in her education analysis as well as Ul-Haq (1995) and Alkire et al. (2009) discuss the importance of school attendance and the need for a learner to effectively engage within class activities (as discussed in section 3.3.2). Alkire et al. discusses a person's need to be able to participate freely and gain a sense of empowerment from their involvement. The final consideration in relation to enjoyment/appreciation is the learner's ability to express themselves freely at the school. In this regard, such concerns were raised by Alkire et al. and Nussbaum (2000) who recognise that the freedom to engage freely amongst colleagues is an important human trait. Where there are efforts to restrict one's expression it generally leads to feelings of depression and fosters poor class participation.

Each of these factors are described in greater detail in Chapter 6 where specific concerns in school literature are highlighted to identify core themes and sub-themes which best describes the make-up of each of these functionings. As noted by Sen in section 3.2 when formulating the Capability Set, the set of identified functionings balance both the materialistic concerns such as financial well-being against non-materialistic concerns such as psychological well-being or a learner's ability to freely express themselves. Furthermore, the Capability Set must recognise the various impediments to access education and identify how these factors could lead learners out of the school system.

3.5. Strengths and Weaknesses of Capability Approach Applications

The identification of a vector of functionings combined with the need to identify the real freedoms available to individuals and communities neatly lends itself to practical application in empirical analysis. Various leading authors have written books that describe how the theory can be adapted and applied. The most notable of these applications is the HDI adopted by the UNDP. Although Sen cautioned against using the theory to make broad statements that generalise patterns across communities and individuals, possibly masking their internal characteristics and challenges, Sen saw the value of identifying measures that informed the debate pertaining to development challenges, arguing that the current economic metrics were far too broad to highlight the core developmental level of a nation.

A known weakness that was faced in developing the HDI was the difficulty in finding comparable data across countries. This difficulty led to various compromises that were made in terms of which indicators were identified to represent the state of international Human Development. Such a difficulty becomes unavoidable when collecting data for every country in the world describing the multi-dimensional state of human development. However, other authors tended to focus on narrower problem spaces or sectors. Capability experts such as Alkire et al. (2009) discuss the importance of measuring freedoms as well as functionings, whilst Education Capability experts such as Robeyns (2011) have focused on the data collection instrument and therefore highlight the difficulties in matching indicators to the theoretical concepts of the approach.

From the above review of the Capability Approach literature to date, one notes that Capability Theory can and has been used to evaluate the education sector. Through the identification of functionings and capabilities it is possible to expand upon the basic set capabilities as identified Sen. This application would allow one to evaluate school attendance and survey the reasoning for learners abandoning such facilities. The identified Capability Set in section 3.4.3 provides the backbone for this study to identify core themes which describe reasons for learners dropping out of the school system.

The experiences from the UNDP who rolled out the Human Development Report, proves that the approach can be applied to extremely complex problem areas. However, there are many concerns that have been highlighted in this implementation which relates to data quality. For example, the manner in which the Human Development Index has been applied has been criticised most notably for data quality and data coverage concerns. Critically, for the development of future indices, the issues of data quality needs to be addressed. Moreover, there is a need to adapt the approach to allow one to find the most appropriate set of indicators applicable to the BRICS and specifically Brazil, India and South Africa. The issues of data quality are addressed in the next chapter.

Chapter 4

Data Quality: From Theory to Practice

In most criticisms of the application of the Capability Approach we find data quality concerns emerge. Within the literature we note that Capability Approach researchers have not attempted to address this matter from a theoretical perspective possibly because the approach is not concerned with data collection and this field is not directly related to human development fields of research. Whilst not theoretically connected, the practicality of selecting indicators in the real world would benefit greatly from a technical data quality assessment that is informed by theoretical aspects of data sciences and more specifically the elements that constitute data quality. Therefore, this study qualifies the search for relevant indicators with an understanding of what data quality is and thereafter identifies how data quality theory can be used to assess the quality of the available public datasets.

In the assessment of data quality that is performed, the key dimensions of data quality are selected based on their relevance to a data user who must consume the data released by the data provider. Whilst this search for data quality dimensions considers the organisational needs, the architectural design and the computational concerns, the selection of dimensions is constrained based on the impact the dimension has on the data user. Furthermore, the outline of the PDQAF is constructed in a manner that can be applied by the data user.

Once the data quality dimensions are identified based on the core tenets of data quality literature, these dimensions are checked for relevance against the priorities of the Brazil, India, South Africa and the International Monetary Fund’s (IMF) Data Quality Assessment Framework (DQAF). From this comparison, relevant dimensions are selected and the steps taken to operationalise data quality assessments are identified and adapted for application to test data sourced from Brazil, India and South Africa.

4.1. Defining Data Quality

It is crucial to test the level of data quality that public institutions have attained in Brazil, India and South Africa when producing datasets. This is to determine the usefulness of the data that is produced for use in policy evaluation or general research. Ben Kiregyera (2015 p 38-41) in his book on the ‘Emerging Data Revolution in Africa’, reviewed the statistical production in

African countries and found what he termed "statistical atrophy" occurred during the 1970s and 1980s. During this period, countries in Africa neglected the development of their statistics due to an under-appreciation of data in policy development. This neglect was coupled with an inadequate provision of resources allocated to such tasks which culminated in the poor quality of data output. Kiregyera termed this the "vicious cycle of statistical under-development" (Kiregyera 2015 p 38). This study determines if this state of under-development of data production continues within Brazil, India and South Africa and therefore undermines the usability of these countries' data outputs.

In reviewing African statistical institutions of the 1990s, Kiregyera (2015, p8 - 11) discussed the international community's shift to results based management. This shift in international management practices has required a change in the manner government programmes are managed and has led to a greater reliance on accurate data. The catchphrase of the 'Managing for Data Results' movement is "Better statistics for better policies and development outcomes" (Kiregyera 2015, p8). Despite this shift, criticisms have been directed towards international indices such as the Human Development Index that are forced to settle for indicators of limited relevance in terms of policy or programme evaluation. These index builders are forced to use indicators that may be of a higher standard of data quality as they are common to the surveyed countries, but may be of little relevance to the analysis. Such practices commonly draw heavy criticism from policy analysts and researchers. However, Kiregyera believes that in recent times developing nations have placed a greater emphasis on the production of data and their control of data quality, thus enabling improved policy evaluation with a broader range of available data.

With the broader range of available data, greater emphasis is placed on assessing the associated data quality. Multiple international organisations have developed assessment tools for data quality and these tools often differ slightly depending on their purpose. The United Nations adopted the UN Fundamental Principles of Official Statistics; the IMF introduced the DQAF. The DQAF was adapted by South Africa when Statistics South Africa introduced the South Africa Statistical Quality Assessment Framework (SASQAF). The UN Fundamental Principles of Official Statistics was developed following the proceedings of Conference of European Statisticians in 1991 (United Nations Statistics Division n.d.) whilst the DQAF was adopted following an Executive Board meeting of the IMF (International Monetary Fund Statistics Department 2003). South Africa's SASQAF is also derived from the principles of the

IMF DQAF. Considering the roots of these international frameworks, it is difficult to determine an originating theory which drives data quality assessment.

Other researchers have suggested other frameworks for testing data or statistics but no underlying theory drives the propagation of these frameworks. Kiregyera (2015, pp 15-19) for example, states “Statistics are good when they have quality (using different dimensions of quality) and integrity, are accessible and are produced efficiently” (Kiregyera 2015, p16). Despite Kiregyera’s elaboration of SMART principles used to assess quality, again no foundational theory is identified detailing what data quality is.

Amongst various organisations, there are differences in what elements constitute good data, good statistics or quality data. Amongst international and national government bodies, it seems the line between data provision and statistical output is blurred without a consistent view on which elements should be assessed. The Health Information and Quality Authority of Ireland (2011), in their assessment of health institutes, referred to a selection of data quality dimensions and noted the best approach was to identify the most common sets of dimensions found in data quality literature. Once such data quality dimensions are identified the assessment process across nations requires a consistent approach in testing data produced by each country using metrics that describe each data quality dimension This approach begs the question of whether the identification of common dimensions is sufficient if not informed by a theoretical grounding, especially if applied across international bodies that have diverse mandates and follow different approaches to managing data quality.

The approach adopted within this study is similar to the Health Information and Quality Authority approach. That is to delve into the literature pertaining to data quality and thereafter identify the core elements of what constitutes data quality and how data quality can be assessed by the users of the data. The key factor here is the emphasis on what the data user can determine based on the type and content of information that is published, that the user can review.

As Kiregyera (2015) notes, there is a shift towards the pursuit of improved data quality. In recognition of this statement, this study tests how this change in data quality consciousness has impacted data pertaining to school dropouts in Brazil, India and South Africa. As Kiregyera states “it is, therefore, important to build a culture of data quality consciousness and to make data quality a cornerstone of statistical work” (Kiregyera 2015, p 137-138). This study examines data produced in these countries by first attempting to fully understand what data quality is and thereafter a Public Data Quality Assessment Framework (PDQAF) is produced premised on this literature.

This study's approach follows that of the Health Information and Quality Authority, however a crucial difference is that data quality is studied from multiple perspectives based on how it is used. Hence, in order to assess whether the data systems providing education related data meet the requirements for quality data, the study of data quality is informed by core texts of Shazia Sadiq (with co-authors, Ge, Helfert, Mcgilvray and Redman (2013)), the editor and co-author of 'Handbook on Data Quality', Martin Eppler's (2006) 'Managing Information Quality' and David Loshin's (2001) 'Enterprise Knowledge Management: The Data Quality Approach'. Sadiq and her fellow authors examine data quality from three positions based on the related requirements of that post in an organisation. Sadiq (Ge et al. 2013) and her co-authors outlines the organisational (management), architectural (system) and computational (statisticians) aspects. Eppler (2006) discusses quality from the perspective of how a user interacts with data and finds it useful. From this perspective Eppler (2006) has produced a framework on quality management based on a series of case studies from the public and private sector. Loshin (2001) recognises the haziness of what data quality is. Although the lay person would assume to know the difference between what good quality and poor quality data is, it is difficult to pin down and assess. Loshin (2001) however focuses on moving away from generalisations and demonstrates how quality can be quantified, measured and improved. A series of supporting authors are consulted per element of data quality to qualify the statements made by these authors.

The following sections unpacks the relevant dimensions of data quality for this study.

4.2. Organisational aspects of data quality

The organisational aspects are derived from an analysis of an organisation owner's perspective to manage data quality concerns holistically within an organisation. These concerns emerge at a global level when an organisation attempts to introduce strategies that involve policies, procedures, roles and standards that cover the broad ambit of data quality across the organisation as a whole (Ge et al. 2013, p ix-x, 1-10). From the various factors which affect the organisation at this global level, the primary factor which affects the organisation as a whole and the external data user, is the manner in which the organisation manages its compilation and provision of metadata. This metadata must describe the efforts taken to manage data quality by the organisation.

Redman in the 'Handbook for Data Quality' (Ge et al. 2013 p15) notes that whilst organisations are improving their data quality processes, there are many that do not recognise the importance of data quality and he believes this lack of appreciation stems from a poor understanding of the contribution quality data makes to an organisation or in this case to the management of school attendance in Brazil, India and South Africa. Managing data quality has various challenges and hence assisting the data user to understand the decisions taken helps them to form a personal assessment of how strong the organisation's data quality processes actually are.

4.2.1. Provision of metadata to the data user

From the data user's perspective when tackling data concepts, the data user needs to examine and work closely with metadata accompanying a published data set. The metadata directly influences the user's perceptions and understanding of what is described in the data. This helps the user to form a sense of surety when using the data, as the user would be able to judge whether the data quality is of a suitable standard. To perform such an assessment, the dataset's metadata is the primary resource used by the data user to determine whether the necessary controls and procedures were put in place to guarantee the data quality in all its dimensions. Therefore, it is crucial to understand what metadata is.

According to the National Information Standards Organization (NISO) (2004) metadata is structured information which describes and makes it easier to generally interact with a particular information resource. It is commonly referred to as data about data. NISO identifies three types of metadata which includes Descriptive Metadata which describes the constitutive elements of a dataset, Structural Metadata which describes how a complex dataset is put together and Administrative Metadata which contains information to help manage and interact with a resource often when the resource is technical in nature.

Franks and Kunde (2006) refer to the ISO Records Management standard which defines metadata as "data describing context, content, and structure of records and their management through time" (Franks and Kunde 2006 p54). They also state that it is important for metadata to be defined and understood in terms of the function that it needs to perform. According to Sidda (2009), when describing the context of the dataset, one must ensure that the information provided is clear and understandable to the data user. Definitions provided need to be clear, whilst ensuring that the language used is not beyond the understanding of the general public.

Therefore, the organisation must determine how to balance using terms of a technical nature to ensure that data users can understand the process and content within the provided dataset.

To summarise, the user's understanding of the provided metadata is crucial in guiding their assessment of the level of data quality that is associated with a particular dataset. This information is used to assess the other dimensions of data quality. It is therefore crucially important that public data providers provide information regarding content such as definitions of data, the context of how data quality is controlled and details pertaining to the structure of the provided dataset.

4.3. Architectural aspects of data quality

The architectural aspects of data quality are identified from the perspective of solution architects who analyse data quality problems arising from experiences when implementing data systems. This view on data quality stems from a technological perspective and experiences in implementing the data quality policies, procedures and standards as specified within the organisation (Ge et al. 2013, p ix-x, 1-10). Hence, these characteristics of data quality are practical in nature and are useful to the data user when assessing a public dataset.

Martin Eppler (2006 p76) provides the rare case where he produces a framework that itemises the core characteristics of data quality but from different perspectives. Firstly, there is the community perspective where data quality is considered from the point of view of one who consumes the data. Secondly, there is the product perspective which refers to data quality elements that describe the soundness of the data output. Thirdly, there is the process perspective which considers issues regarding how the organisation produces the information. As this study evaluates data quality in terms of dimensions which impact the data user, the process perspective is excluded. Lastly, the infrastructure perspective which details the physical constructs required to produce the data and ensure such mechanisms produce quality outputs.

In review of these characteristics, from the data user's point of view, the most crucial perspective is the community level as it highlights the key criteria that users value. In review of the other perspectives, the most valued set of criteria from the product and infrastructure perspectives are selected based on how such criteria impacts the public data user. For purposes of this assessment, where Eppler's criteria have no impact on the data user, they are excluded from this review. The remaining factors that do impact the user are discussed in the following sections.

Quality Perspective	Quality Characteristic
Community Level	Comprehensiveness, Accuracy, Clarity, Applicability
Product Level	Conciseness, Consistency, Currency
Infrastructure Level	Accessibility

Table 11: Information Quality Criteria of Eppler, 2006

4.3.1. Comprehensiveness

The first community level characteristic of data quality is data comprehensiveness. Eppler (2006 p76) finds that the opposite of comprehensiveness is incompleteness. Vannan (2001) identifies comprehensiveness as one of her five core themes of data quality and defines it as the surety that all values necessary within a dataset are present. David Loshin (2001 p103) notes that comprehensiveness is dependent on a user's expectation of what would be included in the dataset. This is consistent with Eppler's view that factors from the community perspective are derived from the user's experience with the data. Loshin further states that data comprehensiveness is achieved when all mandatory and optional attributes are provided within a dataset. This characteristic could refer to a single attribute or be considered collectively when reviewing the dataset as a whole. In assessing how comprehensive a dataset is, Wang and Strong (1996) discuss the breadth, depth and scope of the information required in order to determine whether comprehensiveness has been attained. The authors also highlight that proving that a dataset is comprehensively updated is difficult to determine by those outside of the data collection process (such as public data users).

Redman in the 'Handbook on Data Quality' (Ge et al. 2013 p23) also includes comprehensiveness as one of his four central dimensions of data quality and notes that data users assess comprehensiveness practically in terms of how they are affected and are not aware of the conceptual requirements of data comprehensiveness. Askham, Cook, Doyle and Fereday defined comprehensiveness as the extent to which the dataset has stored data values which meets every potential state as required within the business rules of the organisation. Following this logic we note that comprehensiveness is analysed purely in terms of what the producing organisation deem as complete (Askham et al. 2013). In synthesis of these points, comprehensiveness refers to ensuring that all mandatory and optional components of a dataset are covered ensuring that all business rules of the organisation are met. It is crucial that organisations ensure that they communicate to data users what the internal business rules are which define comprehensiveness.

4.3.2. Accuracy

The second community level characteristic is accuracy. Vannan provides a brief explanation of what data accuracy is and refers to it as data that is free from errors (Vannan 2001). Loshin (2001 p113) has a similar perspective and defines accuracy concisely as the "freedom from defects". Such defects can be described as data values that are inconsistent with an agreed business rule pertaining to a particular data concept. Loshin (2001 p103) also contends that accuracy can be measured by identifying the number of differences between the current incorrect data value and the expected entry. Such a metric, although simple, allows the organisation to internally track accuracy performance over time. This is difficult for a data user to determine if one is unable to determine the expected value.

Eppler (2006 p77) describes accuracy in terms of how it is mentioned in data quality frameworks. Eppler notes that although some frameworks do not explicitly refer the concept of precision or preciseness in place of accuracy. However, despite the difference in terminology, the concept remains a central concept of data quality. Eppler also discusses the trade-off that organisations experience when managing quality as there are times when the need for the most up to date data, outweighs the need for accurate information. In this situation organisations are pressured to release information before testing the accuracy thereof. This is a challenge organisations must be aware of and put in necessary procedures to ensure the data produced meets the requirements for precision.

When assessing accuracy of a dataset it will be important to understand what the expected correct values of the dataset should be and this needs to be compared to the data values that exist in reality. Again where there are expected data values that are resultant from internal business rules, such rules need to be communicated within the organisation's metadata.

4.3.3. Clarity

The next data quality dimension valued by data users, is the manner in which data providers clearly communicate the purpose and contents of the datasets that they release to the public. Loshin's (2001 pp103, 210, 428) definition of clarity stems from how well the terms are defined and understood by data users. Data users experience difficulties when terms are included in a dataset without informing them of their purpose. This applies to tables in a database or dataset, fields within such tables, rules applied to configure the dataset and the metadata that describes the particular dataset informing data users internally and externally to an organisation. Congruently, clarity also applies to the naming conventions that are used within an organisation and are primarily driven by the rules of the organisation.

Clear naming conventions foster improved communication within the organisation and with their public data users. Eppler's (2006 p 79, 83, 88) view of clarity is consistent with Loshin and similarly believes that clarity must be considered from the data user's perspective and therefore the key question that must be asked when evaluating a data release is whether the data is comprehensible to the target audience. Similarly, the Bank of England (Lyon 2008) also listed clarity as one of their core dimensions of quality and noted that its importance to the data user community becomes apparent when they engage the data using the available metadata. The interpretations made by the data users when using the data indicate how clear and comprehensible the released dataset is. This dimension is very valuable within a bank and amongst data providers in general as one of their main concerns relates to limiting the cost of erroneous interpretations of data.

4.3.4. Applicability

The last characteristic within the community perspective is applicability. Eppler (2006 p78) explains that applicability relates to the application and use of a dataset in practice by the data users. The dataset must be assessed to determine if it is useful to data users external to the context of data production. Therefore, the assessment must determine if each data field within a dataset is beneficial and adds value to the user or is superfluous or irrelevant for use. In short, it must be determined if the data released to the public has a practical purpose. Eppler also states that information overload must be avoided when providing data as each included data field or value must be critiqued in terms of what the added benefit a data user gains is from its inclusion in the dataset.

Dasu in the 'Handbook on Data Quality' (Ge et al. 2013 p168), refers to a similar concept but names this 'relevance.' Dasu notes that applicability is determined by assessing what is important for the data user. Dasu compares the data users' needs and as an example, notes that a marketer values very different information when compared to a data engineer. These differences in points of view are consistent for all roles across the data production chain. Therefore, effort must be made to ensure only necessary information pertaining to the data user's role should be exposed to them.

Eppler's discussion of applicability is broader than Dasu who discusses it within the context of the data user. Organisations may find it difficult to tailor their data output to external data users if they are unaware of who their target audience is. However, each field that is exposed to the data user must be assessed to determine if the data user finds value in its inclusion.

4.3.5. Conciseness

The product level refers to the methodological soundness of the data output. Eppler's first product level characteristic is the conciseness of the data output. Eppler (2006 p50, 53, 62) assesses conciseness by determining whether the information provided excludes non-essential and unimportant elements. Conciseness differs from applicability as a concise data output must also ensure precision in how it presents the data. This precision is driven from the level of detail that the data users actually require. Furthermore, if the data is produced in a more concise manner, the task of assessing accuracy becomes easier to undertake. If the data is also presented more concisely, it also supports system efficiency as less system resources are required to transmit or communicate the details of the dataset in question. Eppler argues that an organisation must find the balance between conciseness and comprehensiveness, as to err on the side of conciseness may exclude pertinent information that a user values. Ge and Helfert (2013 p83) notes that a more concise release of data assists the user to better understand the data as it is more to the point. Therefore, removing the superfluous elements and providing detail that is restricted to what interests the user, helps to improve data interpretability.

Whilst conciseness does have an overlap with issues of applicability, a key component relevant for the data user is how well a dataset aggregates data to a necessary level of granularity. Therefore, when assessing public data, it will be important to determine if superfluous data is excluded and if the data is presented at an optimum level of granularity.

4.3.6. Consistency

Eppler's next product level characteristic is data consistency. Loshin (2001 pp110, 111, 216) identifies two forms of consistency, viz., semantic and structural forms. Semantic consistency applies to the rules and definitions applied within a dataset and accompanying metadata. In this regard, where terms are used, they should be applied uniformly throughout the particular dataset and data model. Names and objects should represent a singular entity and ambiguity must be avoided. Such consistency must also be applicable to the different contexts within which the data is used; therefore, the end user and data engineer should hold the same understanding of any particular term. This improves communication throughout an organisation and to their external data users.

Structural consistency refers to the actual representation of data values and its transmission through the data production value chain. Large organisations are in danger of altering data values due to the many copies of tables that propagate through an organisation. Such environments enable inconsistent data to spread and stem from uncoordinated data processes within an organisation (Loshin 2001 pp216, 218, 223). Tejay, Dhillon and Chin (2006) share the structural view of consistency and refer to the representation of a data value across the organisation. Tejay et al. state that consistency is achieved when data values can be tested to determine if the same value has propagated uniformly through the data transmission channels.

Eppler (2006 p50, 52) shares Loshin's distinction of consistency and raises the issue that data can be consistent in format and in value. Whilst the value could pass through the organisation uniformly, if the interpretation of the format and the underlying rules change, the use of the data could be taken out of context.

Assessing consistency of a data element from outside an organisation is difficult as one is not able to refer to the source environment. However, if the applied business rules used to transform a dataset are included in the metadata, the data user will be able to assess whether changes to reported data values follow from the methodology or a possible data fault. The data user can also compare the reported dataset to similar data outputs from the organisation or can be verified against data outputs of other organisations.

4.3.7. Currency and Timeliness

Wand and Wang identify currency as a key dimension of data quality. In this regard, the authors note the close relationship that currency has to timeliness which Eppler also identified within the process level. Wand and Wang (1996) describe currency as the time it takes an organisation to collect, store, process and report a particular data element Eppler (2006 50, 80, 84) suggests the opposing dimension to currency is obsolescence and that the data components processed to produce the released data extract should be up to date. Where data users report on obsolete data, their findings become obsolete too and therefore greater effort should be made to source and release the latest available information.

Loshin (2001 pp xv, 115, 219) makes a distinction timeliness from currency and notes that currency refers to how up to date the information actually is whilst timeliness refers to the expectation one has in receiving a particular extract. This also explains why this particular property is linked to the process level as the timeliness of data receipt is crucial in ensuring the system functions efficiently. Where data delays exist, the data production process slows down. Loshin also notes that timeliness is a variable which can be easily measured by assessing the difference in time in when a dataset was expected versus the point in time it was actually received.

Wand and Wang (1996) note that timeliness can be affected by three factors, viz., the speed at which the system updates its data elements, the frequency of changes mandated by the real-world which the system must react to and the time taken to actually use the data after it is provided. These factors need to be dealt with by organisations to limit the spread of data deficiencies, especially if the organisation strongly values releasing timely and up to date information to the public. Eppler (2006 p84) also notes that in the effort to improve timeliness, organisations also neglect comprehensiveness and accuracy of their data collection efforts. Therefore, an organisation must ensure that the necessary balance between these factors is reached.

On review of these descriptions, timeliness relates to the processing time and the inner workings of a data transactional system. The efficiency of the system bears no impact on the user unless such a factor limits one’s ability to access up to date data. In the public data arena, currency is an apt dimension as the data that is used to describe the school dropout problem needs to be recently published as the trends inferred from the data must provide a current assessment of the situation. Outdated data would be obsolete if it is the only available data that is produced about a particular issue. However, it should be noted that the timeframes regarding these data releases are not real-time updates but rather periodic releases depending on a survey cycle or government data collection strategy.

4.3.8. Accessibility

Eppler’s first infrastructure factor is accessibility. Tejay et al. (2006) note that accessibility implies that data users should find that data is reachable, obtainable and retrievable when needed. Organisations need to pay attention to the channels that they make data available through and ensure such channels are available to their data users. However, as Tejay et al. discuss, data should not be made accessible at the expense of the organisation’s data security requirements. Therefore, organisations should ensure the necessary access rights are instituted. Where processes are put in place to allow only the appropriate data users to gain access to release data, such access procedures should be clearly articulated.

Eppler (2006 pp84, 86, 88) agrees that there is a trade-off between security and accessibility and notes that organisations must guard against the exposure of information which could jeopardise their data security policies. Eppler also notes that when a dataset is evaluated, it must be determined whether the correct people have access to appropriate levels of data based on their respective security profiles. Security can be applied at various levels of granularity within a dataset. Talburt and Zhou in their chapter in the 'Handbook on Data' Quality (Ge et al. 2013 p264) largely agreed with Eppler's perspective on security and noted that most organisations favoured reducing user convenience to counter the risk of exposure. Therefore, greater security procedures are in place to counter the risk.

A key point for organisations is to find the right balance between security, convenience, accessibility and speed. However, for purposes of this study, the emphasis for the public data user is how accessible government institutions have made public data. Where access controls are required, it must be determined if the level of security matches the degree to which the applied security controls are necessary.

In summation, nine of Eppler's Data Quality criteria are relevant when assessing the data quality of publicly produced datasets in Brazil, India and South Africa. From the community perspective all of Eppler's factors have been found to be relevant as they are all factors which directly impact the user and the user's analysis when interacting with the organisation's data output. These include factors of comprehensiveness, accuracy, clarity and applicability. Amongst the product level, three factors affect the user which is how concise the data is when accessed by a data user, how consistently organisations report on such data and how current the data provided is. At the infrastructure level, the accessibility of the data an organisation provides to the public must be assessed. Each of these factors need to be applied and critiqued when used to assess Brazil, India and South Africa's data pertaining to school dropouts.

4.4. Computational aspects of data quality

The computational aspects of data quality emerge when you consider the needs of database experts and statisticians when processing data. These users interact with the data outputs developed by the solutions architect and systems developers. They apply computational techniques to the data that must maintain and promote a suitable level of data quality to ensure the entire data production process is viewed credibly (Ge et al. 2013, p ix-x, 1-10). Whilst their actions are technical in nature, some of their efforts have a direct impact on the data user. Such actions need to be detailed within the released metadata to ensure that the data user understands how such data quality transformations were enacted.

4.4.1. Database integrity

When assessing data quality and specifically data integrity from the data user's perspective, one must consider the documentation made available to the user for them to form an understanding of what data quality constraints have been implemented within the data system. The data user is unable to analyse the internal workings of the system and must make a valued judgement of the data quality constraints using only the metadata. As Arndt and Oman (2007) noted, the metadata provided must be comprehensive to allow users to understand the data's configuration and if any transformations were applied to it.

Leopold Bertossi and Loreto Bravo in their chapter in the 'Handbook on Data Quality' (Ge et al. 2013, p181 - 183) have suggested that there is a need for general data quality assessment methodologies and solutions within organisations. New technologies have emerged that allow for the incorporation of data quality constraints in database tools. These capture data quality issues and provide a means for database developers to identify quality concerns within the data directly and specify a means for adaptive and generic data quality assessment and cleaning mechanisms. Such tools could surface issues such as incorrectly changed data values, duplicate data values or incomplete data values within databases which then prompts users to adopt corrective measures. Where such practices are adopted, the methodology for protecting the organisation's data integrity should be discussed within the dataset's metadata to provide assurances to the data user.

With the emergence of a wide range of database tools, Ceri, Cochrane and Widom (2000) suggest that integrity controls that feature in these database management support tools primarily function as data constraint maintainers. Features such as constraints and database triggers act in response to a violation of a particular rule or if a captured value is found to be beyond the predetermined scope of the entry. The role of these constraints is to enforce integrity controls, however it is also important that such tools are bound by limitations of processing power and are therefore unable to intensively test every low quality scenario due to the trade-off between testing scenarios and processing power.

Fan et al. (2012) in their more recent review of database management systems found that there is a shift to focus more on the possible fixes required to deal with breaches in integrity rather than simply highlighting where data has been found to be erroneous. Typically, systems introduce scenarios or algorithms which detail how the developers and users resolve such inconsistencies. In addition, the authors go further by proposing a selection of particular processes that could be followed to make corrections where inconsistent data is found. Such techniques are possible, due to the large amounts of data that are collected allowing for data imputations with a relative stronger degree of confidence.

Each of these processes described by the authors is important and relevant for the public institutions that produce data as these discuss operational approaches to ensure the integrity of the data system. To promote transparency and ensure that the data users are aware of the constraints and the integrity controls put in place, the business rules of such measures needs to be documented as part of the related dataset's metadata release.

4.4.2. Traceability

A data element is considered traceable when one can determine where the element is derived from. Eppler (2006 p84, 144, 222) notes that traceability is a crucial criterion which assures various data users (from the researcher to the data system engineer) that the source of the information is credible for use. The ability to prove the source of the information and its background through supporting metadata is vital to determine the authenticity of the data. To ensure traceability, organisations provide the author and publisher's information as well as the transformation procedures which are applied to the data. Juran and Godfrey (1998) concur with Eppler's review of traceability and note that some organisations also note the names of all parties that have altered a particular data element within the data processing chain. Therefore, such metadata may provide much greater detail as each party affecting changes is mentioned. Traceability is a key factor when assessing public institutions' data output and the metadata must be reviewed to assess where the data is sourced from and transformed.

Dong, Berti-Equille and Srivastava (2009) conducted a study on source dependence and noted that it is essential to be able to identify conflicting data values from multiple sources and thereafter institute mechanisms to deal with such conflicts. These findings are consistent with Eppler's characteristic of traceability and tend to emanate from the copying of wrong information amongst data users who could act as alternate data providers. It is important to be able to trace the source from where the provided data is extracted as these alternate data providers, who are dependent on the same data source, could easily repeat errors by copying such erroneous data from the original source. In these cases, the original source of the data received must be identified and thereafter the trustworthiness of each dataset that is collected per data source must be assessed.

For the user to assess traceability, data must be published with the corresponding name of the data source. Such information must be included within the metadata of the particular dataset.

The table below summarises the identified data quality dimensions that form the basis of the data quality assessments in this study.

Pillar	**Dimension**
Organisational	Metadata describes the content, context and structure of the dataset
Architectural	Community Factors: Comprehensiveness, Accuracy, Clarity, Applicability
	Product Factors: Conciseness, Consistency, Currency
	Infrastructure Factor: Accessibility
Computational	Integrity, Traceability

Table 12: Selected Data Quality Dimensions for the study

The important next steps required when operationalising these theoretical elements involves identifying whether Brazil, India and South Africa’s data providers recognise these data quality dimensions in their data production process. Where agreement in these factors is found, they are thereafter used to determine the associated quality level of datasets relevant for this study. The identified dimensions presented in the table above are hereon referred to as the Sadiq-Eppler dimensions of data quality.

4.5. Translating Data Quality Dimensions into a Public Data Quality Assessment Framework (PDQAF)

Whilst the Capability Approach theory has clear roots to Amartya Sen who draws from foundational philosophical theories of Aristotle and Rawls, the roots to International Frameworks of Data Quality are less clear. As mentioned earlier in this Chapter, the dimensions of data quality specified in frameworks by the United Nations and the IMF have followed from consultative processes involving representatives of their member countries. Despite the participation of experts in such processes, the relevance of the suggested dimensions of data quality may not be supported by the data quality literature, reviewed in sections 4.2. 4.3 and 4.4.

The chosen approach to developing the Public Data Quality Assessment Framework (PDQAF) is to test the data quality dimensions identified within the literature against those identified by the IMF an in each country involved in the study. Where dimensions are jointly recognised within these frameworks they are adopted for assessment purposes. The literature also assists in identifying any possible weaknesses in the international data quality frameworks.

4.5.1. IMF Data Quality Assessment Framework

The IMF established a data quality assessment framework to provide users of public data the means to make their own quality assessment, which was also called the Data Quality Assessment Framework (DQAF). Although generic in nature, the framework is informed by international best practices, principles and standards of data quality. The framework's foundations emanate from the United Nations Fundamental Principles of Official Statistics and Special Data Dissemination Standards and consist of five dimensions of quality, viz., assurances of integrity; methodological soundness; accuracy and reliability; serviceability and accessibility (International Monetary Fund Statistics Department 2010).

The Prerequisites of Quality dimension refers to the environment from where the data is sourced and involves assessing whether the statistical programme has obtained the necessary resources to function effectively. Although this requirement is raised by the IMF, it does not correspond with any of the data quality dimensions discussed in the previous section. Due to this inconsistency, this dimension is found to be a crucial factor from the data user's perspective and is excluded. Similarly, the Integrity dimension as raised by the IMF is an assessment of the data provider's policies and level of professionalism and is not directly relevant to the data user. This dimension of Integrity differs from the computational perspective of integrity as discussed by Bertossi and Bravo (see section 4.4.1) who referred to the constraints embedded within a data system and not whether the data provider was found to be transparent and ethical. Hence, this dimension is also excluded.

The Methodological Soundness dimension is also quite broad and covers various concepts which were referred to in the Organisational perspective of data quality. The IMF (International Monetary Fund Statistics Department 2010) requires a review of how concepts and definitions are utilized in accordance with international practices, which is consistent with the discussion about the importance of releasing detailed metadata to the public which is inclusive of definitions and transformation methodologies. The assessment of this dimension requires that one identifies whether such practices are conducted in line with international guidelines and standards. The Accuracy and Reliability dimension matches Eppler's description of accuracy and precision and requires two assessments in how closely the data conforms to the identified statistical procedure and how frequently the data output is assessed and verified. Serviceability is consistent with the dimensions of consistency, applicability and timeliness requiring that such principles are applied in the production of a national dataset release. The final dimension Accessibility as highlighted by the IMF refers to the manner that the statistics are publicly presented and whether this is understandable and supported by relevant metadata. These issues correspond with Eppler's discussion on accessibility of data provision as well as the need to make metadata available to the data user.

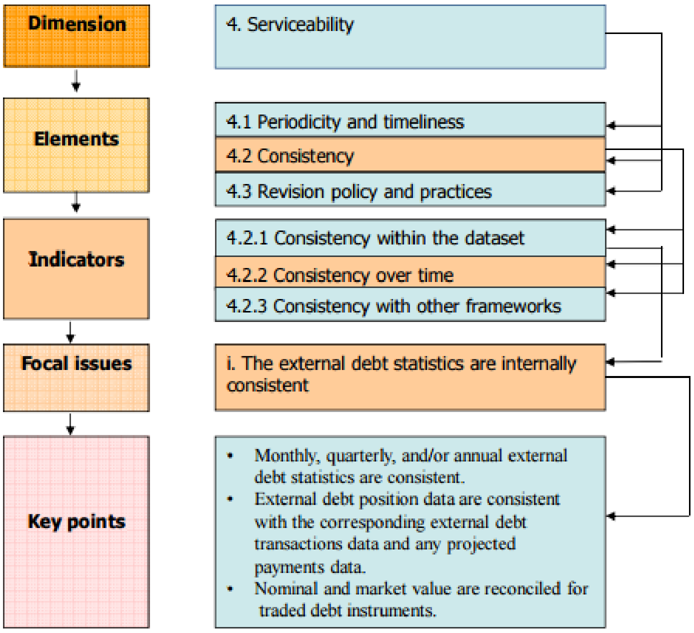
The relation between the IMF dimensions and those discussed in sections 4.2, 4.3 and 4.4 are found in Table 13 below.

IMF Dimension	Data Quality Dimension	Elements
Pre-requisites of quality	Excluded	Legal and institutional environment, Resources, Quality awareness
Integrity	Excluded	Professionalism, Transparency, Ethical Standards
Methodological Soundness	Metadata	Concepts and definitions, Scope, Classification, Basis for recording
Accuracy and reliability	Accuracy	Source data, Statistical techniques, Assessment and validation, Assessment and validation of intermediate data and statistical outputs, Revision studies
Serviceability	Applicability, Currency, Consistency	Relevance, Timeliness and periodicity, Consistency, Revision policy and practice
Accessibility	Accessibility, Metadata	Data accessibility, Metadata accessibility, Assistance to users

Table 13: IMF's Data Quality Assessment Framework's Dimensions and Elements (International Monetary Fund Statistics Department 2010)

A key feature of the DQAF is that each theoretical data quality dimension can be described by specific data quality elements as described within the literature. Thereafter, each element can be further broken into a set of indicators. Carol Carson (2000) in her review of the DQAF noted that the framework and process required an evaluator to identify five levels of detail as described in the framework viz., dimensions, elements, indicators, focal issues and key points as set out in Figure 3. This level of detail is expected to highlight quality concerns that data providers may face. The framework also allows itself to be scaled to an appropriate degree of granularity. This allows one to provide a summary overview of an entire statistical system of a country whilst also providing sufficient detail to describe the system at lower levels.

Furthermore, the framework is generic enough to be applied across countries. The framework also promotes comparability as the language and terms used within the framework tend to inculcate common terminology across diverse countries. Overtime, the use of the framework enables a conversation between data quality evaluators. The framework is also adaptable across a range of subject areas and lends itself towards education needs and other social sector concerns, if needed. Lastly, Carson notes that the process is transparent and reproducible and is therefore more easily accepted amongst governments.


Dimension
Elements
Indicators
Focal issues
Key points
4. Serviceability
4.1 Periodicity and timeliness
4.2 Consistency
4.3 Revision policy and practices
4.2.1 Consistency within the dataset
4.2.2 Consistency over time
4.2.3 Consistency with other frameworks

i. The external debt statistics are internally consistent

- Monthly, quarterly, and/or annual external debt statistics are consistent.
- External debt position data are consistent with the corresponding external debt transactions data and any projected payments data.
- Nominal and market value are reconciled for traded debt instruments.

Figure 3: DQAF Cascading Structure of the IMF DQAF

Laliberte, Grunewald and Probst (2003) compared the IMF's DQAF against Eurostat's quality definition and found that the DQAF approach provided a holistic view of data quality from an international perspective. The authors noted that the strength of the IMF system was the manner it allowed itself to be applied to statistics and the processing thereof. The DQAF also supported the incorporation of data quality benchmarks such as the inclusion of statistical practices which enabled comparisons of quality across countries following a uniform methodology. The authors also emphasised the need to clarify the set of terms which inform the definitions of the data quality definitions, which should be captured in the data's metadata. In Khemangkorn's (1999) review of the application of the DQAF to Thailand's data in the early stages following the publication of the framework, it was found that the DQAF provided an integrated and flexible structure to assess data quality. The study also emphasised the role that the statistical benchmarks played in facilitating the application and practical use of the framework which corresponded with Laliberte et al's findings.

Morten Jerven (2016), in his review of the DQAF cautioned against using the framework as a tool to rank the quality of a country's data as the complexity of assessment would lead to inconclusive results. However, Jerven noted that the major strength of the assessment framework was the manner in which it enabled a common data quality vocabulary across countries, promoting comparison and discussion. The application of the framework lent itself to an adoption of terms and methodologies that were often inconsistent prior to the usage of the framework. Whilst the underlying elements and indicators for each dimension may differ across countries, the methodology of detecting these factors lead to the identification of similar terms that promoted a common conversation when comparing quality dimensions.

The IMF's DQAF, although limited to six data quality dimensions, discusses many common data quality dimensions discussed by Sadiq and Eppler. The strength of the IMF framework actually lies in the methodology of scaling up and down the dimensions to present the core elements that describe data quality. The indicators, focal points and key issues are meant to provide specific measures that describe how one assesses the particular dimension. However, the DQAF does not explicitly state what these items should be. Furthermore, the dimensions such as data lineage tracking, traceability, conciseness and comprehensiveness were not explicitly identified within the DQAF. However, it should be noted that the methodology of the framework does not preclude the expansion of the selected dimensions to cover such concerns.

4.5.2. Data Quality in Brazil

The IMF's DQAF identifies six broad dimensions of data quality, though it may not capture the priorities that each of the countries require. The following section outlines the priorities for data management in Brazil and these priorities are compared to the Sadiq-Eppler dimensions discussed within sections 4.2, 4.3 and 4.4.

Brazil does not follow an official data quality strategy, although its data policies identify data openness and sharing as a priority. The Brazilian Ministry of Planning, Development and Management (Minesterio do Planejamento Desenvolvimento e Gestao 2014) produced the Plan for Open Data which detailed the actions needed to implement an open data strategy. Amongst the various prescriptions of the plan, it was required that geospatial data and its metadata should be disseminated and an e-Government strategy should be established. The plan further notes that the open data strategy should ensure a transparent public administration. Brazil's plan calls for providing access to the most relevant set of data to the broadest cross-section of society with an aim to widen such access and improve the quality of the dataset over time. Furthermore, a minimum set of released metadata was prescribed to accompany publicly released data. Whilst Brazil made various calls for a wider release of public data, the country does not adapt the IMF's DQAF for their purposes.

Germano and Takaoka (2012) discussed the dimensions of Brazil's Government's data quality policy and their move towards an 'Open Data' policy. Open government data refers to the need to expand the availability of government data and information in open formats which are accessible to the public. The included dimensions of data quality in the Open Data plan highlighted by Germano and Takaoka, categorise the dimensions raised by Wang et al (2000) which identified four categories of dimensions, viz., intrinsic, accessibility, contextual and representation. Each of the categories is further described in terms of individual dimensions which can be measured by particular indicators. Many of the dimensions relate to the dimensions mentioned in sections 4.2, 4.3 and 4.4 with a few additional inclusions which are described below in Table 14. The relevant dimensions that correspond to the Sadiq-Eppler dimensions are described below.

Category	**Germano & Takaoko Dimension**	**Relation to Sadiq-Eppler Dimension**	**Brief definition**
Intrinsic	Accuracy	Accuracy	The information is correct and reliable
Accessibility	Accessibility	Accessibility	The information is available and recoverable
Contextual	Relevance	Applicability	The information is pertinent and fulfils a particular purpose
	Timeliness	Currency and Timeliness	The information is regularly updated based on the need it supports
	Integrity/Perfection/Completeness	Integrity, Accuracy, Comprehensiveness	The information meets the required criteria to describe the particular phenomenon
	Appropriate amount	Conciseness	Sufficient detail of the information is provided to meet the purpose of the extract
Representation	Interpretability	Provision of Metadata	The information is supported with relevant definitions,

			language descriptions, symbols and supporting detail such as units
	Ease of understanding	Provision of Metadata	The information is easily understood
	Concise representation	Conciseness	The information is presented in a condensed and easily consumable form
	Consistent representation	Consistency	The range of information provided is made available in the same format

Table 14: Relevant Categories and dimensions of data quality in Brazil (Germano & Takaoka 2012)

Germano and Takaoko (2012) further described eight principles for the open government data initiative to follow. These principles required that the data be complete, primary in nature, current, affordable, process ready, available without request, be under no exclusive control and license free as highlighted by the National Open Government Working Group. These principles for data openness sometimes disagree with the data quality dimensions, where for example, in the case of secure access compared to the need for data to be available without request. The principles of openness are intended to promote innovative use of the data to contribute to the betterment of society in general and to contribute to improved government transparency and ties closely with the dimension of accessibility.

Craveiro, Santana and Albuquerque (2013) developed a data quality assessment framework for government budgetary data in Brazil following the eight open data principles listed earlier. Their study suggested various methods to measure how well the data quality is maintained in terms of each open data principle, by identifying specific metrics that could be used to quantify how a dataset performs per principle. The authors believe their study could be expanded to produce a composite index for measuring the quality of released budgetary data across the various states and municipalities in the country. The authors caution against rushing to publish data if the underlying data quality issues are not addressed.

In Brazil the prioritization of publicising public information is clear, and although the need to publicise data concurs with the principle of accessibility, the Brazilians emphasise the dimension to a greater degree than the other countries, the IMF or amongst data quality authors. A point that is emphasised and also echoes the sentiments of Redman and Sadiq is the need for metadata to clearly express the followed data collection methodologies. The Ministry of

Planning, Development and Management however emphasise the need to a greater extent and require metadata to be provided with a certain standard of coverage. Furthermore, a key concern which must be emphasised amongst the BRICS, as Germano and Takaoko note, in the rush to make greater quantities of data available, the quality of the data production process should not be compromised.

4.5.3. Data Quality in India

Following the review of Brazil's priorities, it is necessary to determine if India also identifies similar priorities as described by the Sadiq-Eppler dimensions of data quality. India's data challenges and policy priorities relate to their highly decentralised statistical system. Responsibilities are shared across the various ministries and spheres of government. Due to the multiple owners and managers of data, there is no single custodian of official statistics in the country which is governed by single data quality standard. The large scale surveys and national accounts of the country are conducted and collected across ministries. The central government coordinates these collections and the Ministry of Statistics and Programme Implementation (MOSPI) (Government of India Ministry of Statistics and Programme Implementation 2009) acts as the nodal agency for such activities which assists with managing statistical collections and implementations. In fulfilling this role, the Ministry of Statistics and Programme Implementation identifies norms and standards pertaining to concepts, definitions, methodologies regarding data collection.

According to the Government of India's National Statistical Commission (2011), the role that the MOSPI plays within the country primarily relates to data management activities such as forming architectures, policies, practices and procedures. The challenges faced by the ministry pertain to (1) the decentralised management of surveys, (2) the need to control the format of the data output, (3) managing non-government produced data, (4) managing the integration of data across vertical and horizontal levels of the government structures and (5) the drawing up of a consistent set of definitions and standards to manage these multitude of concerns. In order to manage these roles, the Ministry promotes the collection, compilation and dissemination of official data. According to the Commission, official data in India should portray the traits needed for quality data such as accuracy, integrity, completeness, validity, consistency, uniformity, density and uniqueness. These traits allow published data to be fit for their intended use and are clearly consistent with the Sadiq-Eppler dimensions identified in this study.

The MOSPI is a subscriber to the IMF's Special Data Dissemination Standard (SDDS) and follows these principles for all data that is published on its site (Government of India Ministry of Statistics and Programme Implementation n.d.). The SDDS provides a guideline for nations regarding how to disseminate their data to the public with respect to four dimensions of dissemination which include the frequency of dissemination, the accessibility of the data, the integrity of the data and quality of the data provided. The SDDS targets datasets and not statistics, and therefore is focused more on data collection activities than numerical interpretation. The SDDS also focuses strongly on the manner in which the data is provided to the public. In this regard, firstly, the documentation describing the methodologies (the metadata) used needs to be shared with the data. Secondly, data comparisons and reconciliations must be conducted to ensure correctness of the data provided. Thirdly, the subscriber to the SDDS must identify when the country deviates from an international best practice in terms of the choice of statistical methodology chosen (International Monetary Fund 2013).

The Indian Department of Science of Technology (2012) has also produced a policy on data sharing within the country and amongst the various pronouncements within the policy. The department also raises the principles that need to be applied to data sharing and the promotion of accessibility. In this manner, the policy calls for Openness, Flexibility, Transparency, Legal Conformity, Protection of Intellectual Property, Formal Responsibility, Professionalism, Standards, Interoperability, Quality, Security, Efficiency, Accountability, Sustainability and Privacy. In addition to the SDDS, one notes that India is party to another data quality framework with a particular set of quality traits. These traits are not entirely exclusive of the Sadiq-Eppler dimensions. However, some are not essential from the data user's perspective.

Indian Agencies	**Quality traits**	**Relation to Sadiq-Eppler**
National Statistical Commission	Accuracy	Accuracy
	Integrity	Integrity
	Completeness	Comprehensiveness
	Validity	Accuracy, Integrity
	Consistency	Consistency
	Uniformity	Consistency
	Density	Conciseness
	Uniqueness	Not a priority for the data user
	Share methodologies in metadata	Importance of Metadata
	Perform Statistical Cross Checks	Accuracy, Integrity, Comprehensiveness

Special Data Dissemination Standard	Document deviations from international best practice	Importance of Metadata, Clarity
Department of Science of Technology	Openness	Accessibility
	Flexibility	Not a priority for the data user
	Transparency	Traceability
	Legal Conformity	Not a priority for the data user
	Protection of Intellectual Property	Not a priority for the data user
	Formal Responsibility	Not a priority for the data user
	Professionalism	Integrity
	Standards	Accuracy, Integrity, Comprehensiveness
	Interoperability	Not a priority for the data user
	Quality	All dimensions relate here
	Security	Not a priority for the data user
	Efficiency	Conciseness
	Accountability	Not a priority for the data user
	Sustainability	Not a priority for the data user
	Privacy	Not a priority for the data user

Table 15: Summary of Quality Dimensions in India

Whilst the National Statistical Commission's traits for quality data closely correspond to the dimensions discussed within the literature highlighted in this study, the Department of Science and Technology in India is very broad and covers issues not relevant to a study of data quality in the country from the data user's perspective. These priorities follow from the country's need to manage a complicated statistical system. Whilst the priority of data quality has been linked to standardising of such practices across the country, there is no single data quality standard. Through the adoption of the IMF SDDS, India closely shares the core data quality dimensions identified by the IMF as well as the Sadiq-Eppler dimensions, however none of these bodies specify how one should go about assessing these dimensions.

4.5.4. South African Statistical Quality Assessment Framework (SASQAF)

In South Africa, Statistics South Africa is the primary collector of official data in the country. Whilst other government departments are responsible for the data collection activities such as surveys and registers, Statistics South Africa administers the Census, various household surveys, financial record keeping and also oversees various government registers. The agency has also produced the South African Statistical Quality Assessment Framework (SASQAF) which was produced in 2008 to provide standards for data collection efforts within the country. The framework provides criteria and procedures for data custodians to evaluate their data collection instruments. The SASQAF is based on a collection of quality indicators and also provides criteria to judge whether the quality of a particular dataset is of a suitable standard to be deemed official statistics. The SASQAF provides a template for which objectives must be whilst the framework does not prescribe how a data collector should go about meeting the particular data quality requirement (Statistics South Africa 2010a). The usability of the framework supports its application across a variety of data collection tools for statistical producers not only within South Africa, as it is targeted to departments and agencies in general. It will need to be determined if an adaption of the framework would be appropriate for use across the BRICS.

The SASQAF is partly derived from the IMF's DQAF as it shares five of the eight quality dimensions. The SASQAF also acts as a tool for the South African Statistician General to evaluate the level of quality attached to a dataset to approve whether the collection meets the requirements for official statistics within the country. The SASQAF assessment thereafter can be used to determine if a specific data collection is assessed as either quality, acceptable, questionable or poor statistics (see Table 16) (Republic of South Africa Government Gazette 2012).

Certification Level	**Description**
Level 4: Quality Statistics	The data collection meets all the quality requirements specified within SASQAF. In meeting these requirements, the statistics drawn from the data collection are 'fit for use' in line with their prescribed mandate.
Level 3: Acceptable Statistics	Such statistics meet most of the quality requirements prescribed within SASQAF. The statistic is deemed 'acceptable' with a proviso highlighting the caution attached to its use. The statistic is still deemed 'fit for use' for the purpose of its design.
Level 2: Questionable Statistics	Questionable statistics meet only a few of the requirements stipulated with the SASQAF. Deductions based on these datasets are limited and is therefore not deemed 'fit for use' in analysis related to its particular collection.
Level 1: Poor Statistics	The lowest quality level, i.e. Poor Statistics is reached when the data collection instrument meets none of the prescribed SASQAF requirements. Such statistics are deemed not fit for use.

Table 16: Structure of the SASQAF (Statistics South Africa 2010b)

The SASQAF is defined in terms of eight quality dimensions, viz., relevance, accuracy, timeliness, accessibility, interpretability, coherence, methodological soundness and integrity. For a SASQAF assessment to take place the data collection instrument must also meet the prerequisite for quality which focuses on the institutional and organisational conditions which enable data quality. These prerequisites are similar to those required by the IMF in section 4.5.1 (Statistics South Africa 2008). As noted in Table 17 many of these quality dimensions concur with those raised within sections 4.2, 4.3 and 4.4. However, the strength of the SASQAF is found in the details of each key components which explains how each dimension can be assessed. This detail was not provided by the IMF DQAF or within Brazil and India's data priorities.

Despite South Africa's detailed data quality framework, however, it has been noted that the introduction of quality data still suffers in the implementation phase. Mukwevho and Jacobs (2012) discussed the importance of having a quality electronic records management system in government departments. Without naming the department the authors analysed and highlighted a single National Department's reluctance to embrace the SASQAF. It was noted that the culture of the department led to a lack of will to adopt the SASQAF and the requirements detailed within. In this department it was found that despite the spirit and mission of government to shift towards SASQAF, the actual political will and change management processes required to do so were lacking. This leads to poor record keeping, reporting and a loss of vital information needed for decision-making and ultimately undermines social upliftment.

Within South Africa many government departments have not adopted the SASQAF. Perhaps they do not recognise the value that the change represents. The Statistician General of South Africa in 2001 noted some of the challenges that face the country and stated that the country requires a wide array of quantitative data for decision making and programme implementation. Lehohla (2001) further noted that the country ran the severe risk of basing decisions on data with poor quality. In 2005 in a presentation also made by the Statistician General, Lehohla (2005) noted that official statistics in the country had five critical functions, viz., promoting social debate, ensuring adequate resource allocation where required, assisting the development of national interventions, enabling the monitoring of government progress and ensuring an official feedback process in relation to government programmes. Statistics South Africa's Strategic Plan of 2010/11 – 2014/15 (Statistics South Africa 2010c) emphasised the SASQAF in South Africa's government and noted fragile IT systems as well as a need to limit bureaucracy as factors which hamper the full implementation of the framework. Although the framework was still recognised as the key enabler in addressing the data quality gap which affected South Africa.

Dimension	Relation to Sadiq-Eppler Dimension	Key Component
Prerequisite of Quality	Not relevant to the data user	• Legal and institutional environment including Memoranda of Understanding (MoUs) or Service Level Agreements (SLAs) • Privacy and confidentiality • Resources are commensurate with the needs of statistical programmes • Quality is the cornerstone of statistical work
Relevance	Applicability	• Why do you need to conduct a survey or collect data? • Who are the users of the statistics? • What are their known needs? • How well does the output meet these needs? • Are user needs monitored and fed back into the design process?
Accuracy	Accuracy, Currency, Comprehensiveness	• Assessment of sampling errors where sampling was used • Assessment of coverage of data collection in comparison to the target population • Assessment of response rates and estimates of the impact of imputation • Assessment of non-sampling errors and any other serious accuracy or consistency problems with the survey results • Data capture errors • Source data available provide an adequate basis to compile statistics. (e.g. administrative records) • Source data reasonably approximate the definitions, scope, classifications, valuation, and time of recording required • Source data are timely
Timeliness	Currency	• Production time (for the entire survey) • Frequency of release • Punctuality of release
Accessibility	Accessibility	• Catalogue systems are available in the organ of state or statistical agency • Delivery systems to access information • Information and metadata coverage is adequate • Measure of catalogue and delivery systems performance • Presentation of statistics in a meaningful way • Means of sharing data between stakeholders
Interpretability	Clarity, Available Metadata	• Concepts, definitions and classifications underlying the data • Metadata on the methodology used to collect and compile the data

Dimension	Relation to Sadiq-Eppler Dimension	Key Component
Coherence	Clarity, Consistency	• The use of common concepts within and between series • Common definitions within and between series • Common variables and classifications within and between statistical series • The use of common methodologies and systems for data collection and processing within series • Use of common methodology for various processing steps of a survey such as edits and imputations within series
Methodological soundness	Traceability, Data Lineage Tracking	• International norms and standards on methods • Data compilation methods employ acceptable procedures • Other statistical procedures employ sound statistical techniques • Revision policy, transparent, and those studies of revisions are done and made public
Integrity	Integrity	• Professionalism and ethical standards in guiding policies and practices, which should be reinforced by their transparency standards • Assurances that statistics are produced on an impartial basis • Ethical standards are guided by policies and procedures

Table 17: Prerequisites and Quality Dimensions of the SASQAF – Key Components quoted directly from Statistics South Africa 2008

In South Africa, whilst there is a single framework and approach to identifying which datasets constitute official data, many government departments and statistical agents do not follow the SASQAF. Furthermore, it is also noted that following the IMF DQAF Brazil and India have not specified in detail what indicators and steps should be followed to assess data quality. Therefore, in developing the data quality framework for this study the SASQAF will be strongly referred to when detailing steps to unpack the specific dimensions described by Sadiq and Eppler amongst other authors.

4.6. Outline of the Public Data Quality Assessment Framework (PDQAF) in Brazil, India and South Africa

The selected public data quality assessment is based on a combination of structures by researchers such as Sadiq (Ge et al. 2013), Eppler (2006), other data quality authors, the IMF (2010) Data Quality Assessment Framework, priorities from India and Brazil as well as South Africa’s Statistical Quality Assessment Framework (Statistics South Africa 2008) as described in detail in the previous sections. The deviations from these frameworks is due to the study's focus on the requirements of the data user. This assessment is not to critique the data producing organisation and their embedded data quality rules but rather assess the final data product that the data user is able to interact with. In summary, the factors that are applicable to these three countries are detailed in Table 18.

Public Data Quality Assessment Dimensions	**Brazil**	**India**	**South Africa**
1. Organisational Dimensions of Data Quality			
a. Assess the usefulness of the Metadata			
i. Content of dataset	Yes	Yes	Yes
ii. Context of dataset	Yes		Yes
iii. Structure of dataset	Yes	Yes	Yes
2. Architectural Dimensions of Data Quality:			
a. At Community Level in terms of:			
i. Comprehensiveness	Yes	Yes	Yes
ii. Accuracy	Yes	Yes	Yes
iii. Clarity		Yes	Yes
iv. Applicability	Yes		Yes
b. At the Product Level in terms of:			
i. Conciseness	Yes	Yes	
ii. Consistency	Yes	Yes	Yes
iii. Currency	Yes		Yes
d. At the Infrastructure level in terms of data			
i. Accessibility	Yes	Yes	Yes
3. Identify Computational Dimensions of Data Quality in terms of:			
a. Data Integrity	Yes	Yes	Yes
b. Traceability		Yes	Yes

Table 18: Data Quality Dimensions Applied to Brazil, India and South Africa

The IMF's DQAF provides the essential process flow for unpacking how to assess each dimension. Hence, for each dimension listed in Table 18, the following process must be followed:

i. Identify relevant elements that constitute each dimension

ii. Examine each element to identify the particular indicators which best describe the performance of each element

iii. For each indicator, information must be found within the dataset that describes the focal issues and key points related to the indicator.

For each of the eight dimensions defined within South Africa, the SASQAF provides descriptions, key indicators and standards to assess each indicator. This framework can be adapted to fit the PDQAF model. Relevant indicators will be selected from the framework, where there are gaps, such as in the Conciseness dimension, supporting indicators will be determined by consulting the literature for additional details. A key difference between this framework and the IMF DQAF is the approach to convert the key points into either positive or negative traits which are derived from the assessment levels specified in the SASQAF for each dimension. In this manner, the Framework will immediately allow one to determine whether a dataset displays positive or negative data quality characteristics.

To explain how the PDQAF and the Capability theory are integrated, the following section, the Empirical Framework, discusses the methodology for this study with an emphasis on how these dissimilar theories can be jointly applied to identify relevant indicators to describe the school dropout phenomenon.

Part II

Empirical Framework

Chapter 5

Methodology

In order to determine whether locally produced public data from BRICS can be used to compile indicators describing the social concerns of the bloc, a deep assessment of a particular social phenomenon affecting specifically Brazil, India and South Africa is conducted. Using the example of a complex problem such as the school dropout phenomenon which affects each of the selected nations in a different manner, allows one to determine if the identified approach involving a capability assessment and data quality assessment is practical to follow when expanded across the broad base of social indicators as needed by the BRICS member states.

To assess this complex multidimensional phenomenon of school dropouts, which notable authors across the countries argue is due to the cumulative effect of factors faced by the learner over the length of their attendance of school, Amartya Sen's Capability Approach is selected as most appropriate to express the pluralistic nature of this phenomenon. This approach is highly apt when examining the social sector as the theory has already been applied within the education sector in numerous studies.

The manner in which the Capability Approach is operationalised follows the steps itemised by Flavio Comim, a member of the Human Development and Capability Association, an organisation which specialises in Capability Approach assessments. This involves assessing the target environment for the study and thereafter identifying the real freedoms that are accessible by data users and the makeup of a Capability Set which is comprised of a selection of functionings which the learners aspire to attain (see section 3.1). To identify these components as valued by the learners, current literature on school dropouts in Brazil, India, South Africa and internationally was reviewed to identify which factors can be incorporated into this study.

This approach of producing a 'wish list' of possible factors to assess maybe theoretically well founded but could be practically difficult to implement. The adoption of the second theory in this study introduces the element of practicality. the dimensions of data quality which form the basis of the data quality assessment was elicited from data quality literature produced by Shazia Sadiq, Martin Eppler and others. Thereafter, data collection instruments are selected if they are consistent with the selected set of freedoms and functionings. These datasets are assessed in terms of the identified dimensions of data quality but, limited to a selection of dimensions which can only be assessed from the perspective of a regular data user. This grounding of the study, in terms of practical and available data, can then be assessed using this study's framework titled as the Public Data Quality Assessment Framework (PDQAF), which is largely modelled on frameworks produced by the International Monetary Fund and Statistics South Africa.

The joint application of the two assessment frameworks brings together two diverse fields united from the perspective of a data user, who requires relevant data which quantifies the complex and multidimensional nature of learners who drop out of school. This data needs to be trustworthy to allow members of the BRICS the ability formulate reliable and independent analysis informed by local information. This study is premised on a belief that the application of the Capability Approach in conjunction with a data quality assessment framework allows for the systematic identification and assessment of indicators within Brazil, India and South Africa. This premise was tested against public data available from these countries following the workflow described in Figure 4.

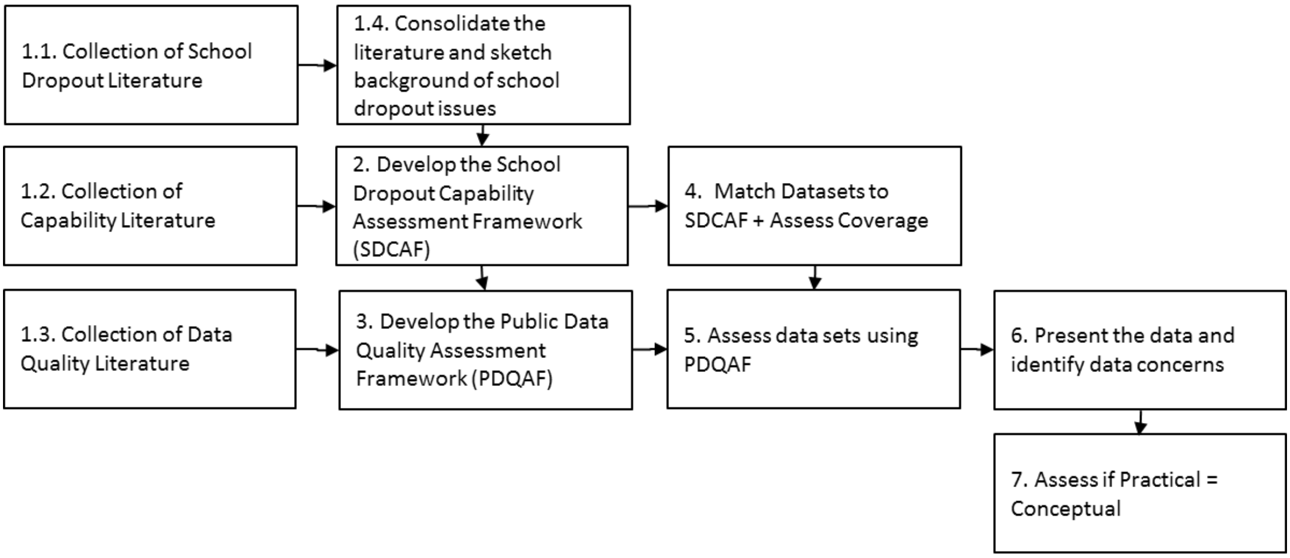

1.1. Collection of School Dropout Literature
1.4. Consolidate the literature and sketch background of school dropout issues
1.2. Collection of Capability Literature
2. Develop the School Dropout Capability Assessment Framework (SDCAF)
4. Match Datasets to SDCAF + Assess Coverage
1.3. Collection of Data Quality Literature
3. Develop the Public Data Quality Assessment Framework (PDQAF)
5. Assess data sets using PDQAF
6. Present the data and identify data concerns
7. Assess if Practical = Conceptual

Figure 4: Workflow sequence for the study

5.1. Collect Literature

In keeping with the manner in which empirical studies are conducted, the first step of the project as described in Figure 4 above, involves collecting data for the study. This includes firstly collecting reports and journal articles from the three countries pertaining to this study that details the primary concerns related to school dropout issues. Each country faces challenges that inform the priorities of the country and their differences in perspective. These challenges describe the background from which this study emanates and thereafter must be analysed in terms of literature pertaining to Capability Approach described by Amartya Sen and other notable authors who have applied the theory, taking specific recognition of applications within the education field. In addition to capability literature, information related to the definition of data quality and its management thereof was collected. The central data quality study which guides the assessment, was edited by Shazia Sadiq who produced 'The Handbook on Data Quality' (Ge et al. 2013). It provides the foundation for the study to conceptualise what data quality is and how it can be assessed by an external data user. Furthermore, the actual publicly available datasets from Brazil, India and South Africa have been sourced.

The literature required for the study is categorised in Table 19, provided below, which identifies the focus area for each component of the study. The section on introducing the issues related to school dropouts in Brazil, India and South Africa discusses the background of these issues, specific to the context of the country. Inputs from Pritchett (2004) provide the central thread for the global context of the phenomena. In addition, the factors that drive school dropouts and the main concerns which pull learners out of school are identified. Following this, the dropout rates used within the three countries are identified. The formulae presented discuss the limitation of using a single incomparable measure across the three countries to represent a complex and multi-dimensional practice. As the factors affecting each country are germane to the country alone, these factors are identified individually. Studies by Spaull (2013, 2015), Gustafsson (2011), Fleisch et al. (2012), Branson et al. (2014), (amongst others) are reviewed. The South African phenomenon, identified by these authors, are summarised in Table 4 in Chapter 2. Issues such as financial limitations, the quality of learning, lack of access to basic services and overcrowding are discussed and each have an impact on learners leaving school. In India Chugh (2011), Gouda and Sekher (2014) and Kumar et al. (2013) are reviewed (see Chapter 2, Table 5) and these authors highlight similar issues as in South Africa with the addition of Gender, Caste, Tribe and Religious disparities. In Brazil, the concerns relate to the opportunity cost of education that families contend with (see Chapter 2, Table 6). Often school attendance is neglected due to the need of a learner to work and contribute financially to the household.

The Capability Approach is primarily informed by the works of Amartya Sen. Firstly; there is a need to unpack the various concepts discussed within Capability literature. In this regard papers and books by Sen (1983, 1985, 1987, 1992, 1999), Alkire et al. (2008, 2009), Robeyns (2011) and Comim et al. (2008) are important to understand the terms that are used within various scenarios. Whilst Sen's discussion tends to follow philosophical lines regarding the motivating arguments for including the particular term within the theory, the other authors discuss how the approach can be practically applied. Sen's theory has developed over the years. His books from 1980 including 'Equality of what' 'Development: Which way now?' (1983), 'Commodities and Capabilities' (1985), 'The Standard of Living' (1987), 'Inequality Re-examined' (1992) and 'Development as Freedom' (1999). Each detail minor changes to Sen's position on the concepts of Development, Capability and Freedom. In addition to understanding Sen's changing understanding, it is necessary to review how the works are received by notable authors within the field such as Clark (2005), Nussbaum (1993, 2000, 2010), Streeten (2000) and others. The approach has been applied most notably by the UNDP (1990) within the Human Development Index. These implementations further need to be reviewed to understand the principles and approaches adopted to ensure successful implementations. Lastly the application of the approach in the education sector needs to be examined. In this respect discussions provided by Walker (2005), Unterhalter (2009) and Unterhalter et al. (2015) are pivotal and are supported by Saito (2003) and Vermeulen (2015).

In exploring the Capability Approach and basic factors affecting school dropouts, it becomes clear that numerous datasets are relevant for reporting on school dropouts. However, the data selected must be filtered based on the degree of quality attained in its production. Kiregyera (2015) discusses how public data quality is beginning to improve in Africa and identifies various frameworks for assessing quality. However, none of the frameworks provide a comprehensive view of what data quality means. The work of Sadiq et al. (2013) clarifies the three pillars which inform data quality and relate to its perspective of use. To best understand data quality, quality issues must be considered at an organisational, architectural and computational level. The works of Martin Eppler (2006) and David Loshin (2001) are instrumental in helping to clarify the dimensions of data quality which are referred to at each level. Various other authors attest to how data quality impacts their organisations and can best be managed. Their works are reviewed and compared to the assertions made by the authors within the 'Handbook for Data Quality'.

Section	**Focus Area**
School Dropout Phenomenon	Background to the phenomenon
	Driving Factors of School Dropout
	Current School Dropout Rate Definition Used
	Dropout Factors in SA

Section	Focus Area
	Dropout Factors in India
	Dropout Factors in Brazil
Capability Approach	Introduction of the Capability Approach
	The Progression of the Capability Approach
	Application of the Capability Approach
	The Capability Approach Applied to Education Sector
Data Quality	Results Based Management Approach
	Organisational Aspects of Data Quality
	Architectural Aspects of Data Quality
	Computational Aspects of Data
	Data Quality in Practice

Table 19: Section structure of relevant literature

5.2. Develop the School Dropout Capability Assessment Framework

The literature collected in step 1 of the study (specifically the literature describing the school dropout phenomenon in the three countries and internationally) is used, together with the theory of the Capability Approach as discussed by Sen and operationalised by various other authors who have operationalised the theory in previous assessments and within the education space as well. The Capability Approach provides the backbone for the framework as it assists in developing the structure in terms of determining the level of real freedom attained by learners contrasted against the collection of functionings which describe the valued aspirations of learners.

Informed by the school dropout literature, a central real freedom (or capability) related to learners who drop out of school, is selected. This capability forms the central pillar to which the other themes discussed within the literature are thereafter associated. In this manner the themes found in the literature are organised to clarify the relationships between concepts and also identify which sets of information fall beyond the scope of the central theme. Following this process, the School Dropout Capability Assessment Framework (SDCAF) is developed based on the central capability and the ideas that emerge from the literature.

Using the Capability Approach theory, an effort is made to identify and structure a Capability Set which is inclusive of various functionings that underpin the school dropout phenomenon in relation to the priorities identified within the literature. The direct approach as discussed by Sen is used to go about selecting the capabilities and functionings (see section 3.1) (Sen 1999, p81-83). Whilst Sen’s theories are individualistic, the theory takes into account the collective experiences and desires of the three nations involved. Thereafter, for each functioning within the Capability Set, the key themes and sub-themes across the 3 countries are identified. The thematic descriptions provide the link between the theoretical concept and the practical indicator.

The framework outlines broad sub-themes which describes states that are valued by the learners. Thereafter, datasets are linked against the sub-theme, where a particular data collection instrument contains relevant questions or variables. As the objective of developing the framework is to select pertinent datasets, this process is crucial in answering the first research question of this study regarding the identification of possible datasets to best describe the school dropout phenomena (NB. the selection of datasets is qualified through the application of the Public Data Quality Assessment Framework discussed in step 3). Also note that the construction of this framework (SDCAF) directly answers the second research question of the study pertaining to what frameworks can be used to assess school dropouts.

A major challenge that lies in operationalising the Capability Approach is the theoretical under specification of how the approach should be applied by its primary author, Amartya Sen. Sen's intention was not for the approach to be prescriptive in its use but rather be context dependent. Whilst the lack of specification from Sen is apparent, he attaches great importance to its practical application. Sen states "the approach must nevertheless be practical in the sense of being usable for actual assessments of the living standard" (Sen et al. 1987 p20). Sen raises this point when he argues that the application of the approach must balance relevance of the approach against its usability as an approach that is too complex actually loses its relevance. The fact that Sen was not prescriptive in how the approach should be applied does not preclude one from producing a strategy to practically apply the theoretical requirements. In doing so, one must be careful to ensure that the various concepts of freedoms, capability, functionings and agency are introduced in manner consistent with Sen's intentions.

Comim et al. (2008) attempts to address Sen's concerns with the tasks he proposes to operationalise the Capability Approach. This study adopts a selection of Comim's recommended tasks as the approach provides a simple process flow for conducting such a study, whilst also providing guidance on how to introduce the various relevant concepts that drive such a study. Comim's process addresses how an evaluative framework could be produced and provides an example of its application. This is done by attempting to balance the conceptual against the practical concerns and by identifying the most pressing concerns related to the field of study to elicit and evaluate available datasets. Such datasets can be ranked based on their alignment to the most pressing concerns. Furthermore, trade-offs will need to be made based on the practical limitations of available data. Once indicators are selected, analysed, and assessed in terms of data quality, the conceptual requirements are compared against the practical reality to assess whether the conceptual values are captured and reflected accurately.

Task	Method
1. Examine the environment	The first task outlined by Comim et al. is to collect information on the communities that were to be studied. This involves meetings with community members, a review of personal articles as well as literature collection and review that outlines the most pressing problems were that required attention. Once information about the most pressing concerns is identified, they are then categorised to reflect the priority concerns in relation to the Capability Approach.
2. Structuring the data collection activities	Following the identification of the most pressing concerns, the findings of that exercise inform the manner in which a questionnaire is structured. Generally, this would require socio-economic variables, subjective variables, functionings and capabilities regarding individuals and also social functionings and capabilities. In cases where the researcher is not conducting a survey a similar identification of secondary data is conducted based on datasets that contain variables relevant to the study. For the purposes of this study, the key data providers from each of the countries will be examined to find what data is available that can be used to inform the selected functionings.
3. Ranking the themes	The selection of datasets follows the structuring and categorisation of the most pressing needs of the community. During this process it is important to combine subjective and objective data measures. Objective measures help define the nature of the phenomenon whilst subjective datasets inform the context pertaining to the study. When assessing the current performance of a geographic region, the multidimensionality of the region must be considered when performing a multivariate aggregation. In this regard, a ranking exercise is required to identify which variables require weightings which are used in the construction of a composite index. However, for purposes of this study, a composite index is not produced.
4. Choosing the datasets	Following the ranking of themes in a manner compliant with the conceptual and practical needs of the Capability Approach, one must then make a selection of variables from dataset that best describes the informational space under examination. Comim et al. recommends that the informational space is categorised by resource, subjective and capability related indicators. Greater exploration of the quality of the data is conducted in the second phase of the study where data quality is examined more deeply.

Table 20: Tasks in operationalising the Capability Approach (Comim et al. 2008 p187 - 197)

By linking datasets to our identification of the learner's valued state of beings and doings, the study is able to link the conceptual to the practical. Such a linkage expresses the selection of Capabilities and related functionings. It also expresses what Brazil, India and South Africa find as a real opportunity or a desired state of attainment. Whilst Comim et al. (2008) and other authors note the difficulty associated with identifying truly objective measures that capture these desired states or real freedoms, there is value in expressing the available array of data instead of focusing on a simplistic learner dropout rate or commodity based perspectives that traditional economists employ.

As part of tasks 1 and 2 involved in operationalising the Capability Approach the Capability Set as described in section 3.4.3 requires expanding the identified set of functionings which were identified in section 3.4.3. Using the literature outlined in Chapter 2, the various factors which are identified by school dropout experts are associated with the functionings identified by Capability Approach experts. Therefore, as part of examining the environment, the various studies that have been discussed are referred to and are critically assessed in terms of their relations to these functionings. Furthermore, the specific factors related to the community, school, household, individual and demographic contexts that are identified in Table 8 in Chapter 2 are associated with the appropriate functioning. Using these factors, the themes and sub-themes which describe these broad concerns are defined. Following the identification of sub-themes, data providers from Brazil, India and South Africa are reviewed and data that are related to these concerns are sourced. The details of this step are discussed in section 5.4.

5.3. Develop a Public Data Quality Assessment Framework

While the need to represent the community's set of real freedoms based on their socio-economic circumstances and their desired states of beings and doings is important, it is also equally important to select data of a higher data quality standard, where possible. For this study, the most appropriate manner of making this selection involves combining data quality theories which are used to produce the Public Data Quality Assessment Framework (PDQAF). As discussed Chapter 4, the PDQAF is based on a combination of frameworks and literature, such as the IMF's DQAF, Statistics South Africa's SASQAF and the selected Sadiq-Eppler data quality dimensions.

The IMF's (2010) DQAF provides the cascading structure for the PDQAF and identifies how one can disaggregate the selected Sadiq-Eppler dimensions into its constituent elements, indicators and specific key points. South Africa's SASQAF, provides the means to identify positive and negative key points which can be used to determine whether the particular dimension, on aggregate displays either good or poor data quality traits.

Data Quality Dimensions			
1. Assess Data Quality performance per dataset in terms of the following data quality dimensions: a. Organisational perspective of Data Quality i. Assess if the Metadata describes 1. Content of dataset 2. Context of dataset 3. Structure of dataset b. Architectural perspective of Data Quality: i. At Community Level in terms of: 1. Comprehensiveness 2. Accuracy 3. Clarity 4. Applicability ii. At the Product Level in terms of: 1. Conciseness 2. Consistency 3. Currency iii. At the Infrastructure level in terms of data 1. Accessibility c. Computational perspective of Data Quality in terms of: i. Data Integrity **ii.** Traceability	Identify Elements	Identify Indicators	Identify Focal Issues and Positive/Negative Data Quality Traits

Table 21: Data quality Assessment Framework

5.4. Identify relevant datasets as per thematic areas of the School Dropout Capability Assessment Framework

The SDCAF identifies the core capabilities, functionings and thematic areas relevant for application. These thematic areas are used to identify questions within each country's specific datasets. The application of the SDCAF helps answer research question three of the study, as the selection process is paramount in deciding which datasets are useful within the school attendance/school dropout space.

Data providers for each of the three countries are identified in the next section. These organisations provide public data in Brazil, India and South Africa. Data produced by these organisations are assessed if they are fit for use via the SDCAF. Once the data is collected, it is assessed using the PDQAF to determine if it sufficiently meets the needs for suitable data quality.

5.4.1. Brazil's basic education data providers

A review of data sources in Brazil has identified four institutions that collect and disseminate relevant data for this study in Brazil, viz. the Brazilian Institute for Geography and Statistics/Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira de Geografia

Estatística (IBGE), the Brazilian Ministry of Education/Ministério da Educação, the National Institute of Educational Studies and Research/Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) and the National System of Public Security Information, Prison and on Drug. The obvious and most prominent challenge in sourcing data and information from Brazil is the language barrier. Data, documentation and general information is often only published in the Brazilian variant of Portuguese. This challenge is overcome by using online language translation tools provided by Google and Microsoft to provide the content in English. Whilst the grammar translation is often not perfect, one is able to elicit the context and discussion flow after closer examination of the translated documents.

The Instituto Brasileiro de Geographia e Estatistica (IBGE) (Brazilian Institute of Geography and Statistics) (Instituto Brasileiro de Geografia e Estatística n.d.) is the premier statistical institute in Brazil and is responsible for the Brazilian National Census, the many household surveys and various other data collection processes in the country. Fortunately, some of the IBGE information is published in English. The IBGE offers a statistical portal on their website which provides access to data at various levels of geography as well as supporting metadata.

The Ministério da Educação (Ministry of Education) manages the various programs for Education in the country which includes programs supporting basic education, literacy, continuing education and diversity. The Ministry also provides access to a portal called Painel de Controle do MEC (MEC Control Panel) where statistics are disseminated. This includes data regarding the trajectory of students, their socio economic context and their results from participation in the countries national examinations (Brazil Ministry of Education n.d.).

The Instituto Nacional de Estudos e Pesquisas (INEP) Educacionais Anísio Teixeira (National Institute of Educational Studies and Research in English) is responsible for the assessment of the functioning of the Ministry of Education. The institute, established in 1990; collects, processes and analyses education data from across the country. INEP is responsible for major school surveys such as the annual School Census which provides granular school information. Basic education indicators are available from the their portal (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira n.d.).

The Crime Statistics of Brazil are released online by the Sistema Nacional de Informações de Segurança Pública (SINESP) (National System of Public Security Information, Prison and on Drug in English). The data made available on the Crime Statistics Portal includes time series information regarding robberies, homicides, rapes and injuries leading to death amongst other crime trends which can be useful to use as a proxy for how safe learners may feel in their respective communities (Sistema Nacional de Informações de Segurança Pública (SINESP) n.d.).

5.4.2. India's basic education data providers

Publicly available data produced by the government of India for the education sector is available from the Office of the Registrar General and Census Commission, the District Information System for Education and the Ministry of Human Resource Development. The primary challenge when sourcing data from India is in understanding the roles and responsibilities of the complex Indian government structure which is comprised of 54 ministries that manage 29 states. The key body responsible for Education in the country is the Ministry of Human Resource Development.

The Office of the Registrar General and Census Commission is housed within the Ministry of Home Affairs and is responsible for the roll out of the national census and has the task of enumerating India's massive population. The census provides basic socio-economic data about the population whilst also providing data on progress in the education sector(Government of India Ministry of Home Affairs Office of the Registrar General & Census Commissioner 2011).

The District Information System for Education (DISE) was established in 1994 to assist the "educationally backward districts of India" (District Information System for Education (DISE) n.d.). DISE is developing an information base to support planning and monitoring projects and is required to overcome the data challenges experienced by the Ministry of Human Resource Development. DISE collects data relevant to Elementary education in the country and collects data that describes demographics, access and participation, the learning environment, human resources, financing, student achievement as well as efficiency and effectiveness indicators.

The Ministry of Human Resource Development is responsible for Education in India. The ministry includes two departments viz. the Department of School Education and Literacy as well as the Department of Higher Education. The School Education and Literacy department collects data about basic education and focusses on the universalization of education in India. The Ministry manages a statistics portal which disseminates the followings sets of information to the public: examination results, education expenditure, Indian standard classification of education, national level education statistics, population projects and school level education statistics (Government of India Ministry of Human Resource Development n.d.).

The Indian Ministry of Statistics and Programme Implementation was established in 1999. The ministry is concerned with data coverage and quality needs related to statistic production. The ministry releases socio-economic data sourced from administrative sources, surveys and censuses. Some of the sources are also non-official studies, whilst the surveys are based on scientific sampling methods (Government of India Ministry of Statistics and Programme Implementation n.d.). Data is also provided via the Open Government Data (OGD) Platform India which includes education indicators as well as population data based on various National Sample Surveys and other sources. The portal is used as a means to publish datasets, documents, services, tools and applications collected by the national ministries which were intended for public use (Open Government Data Platform India n.d.).

The Demographic and Health Survey Programme also runs the Demographic and Health Survey in India and is managed by Inner City Fund (ICF) International with funding provided by the United States Agency for International Development (USAID). The survey is conducted internationally with various implementations conducted across countries over many years. Similar iterations of the survey were also conducted within South Africa in 1998, 2003 and currently in 2016 as well. In India the survey was conducted in 2005/06 and is underway in India in 2015/16. The data collected covers aspects of demographics and health issues. The Health issues relate to communicable and non-communicable diseases as well as risky behaviour and nutritional concerns amongst other notable themes, which will be useful in scanning for data pertaining to the learner’s wanting to physically and mentally healthy (Demographic and Health Survey Programme n.d.).

5.4.3. South Africa’s basic education data providers

Data in South Africa that is relevant for this study is available from the Department of Basic Education, Statistics South Africa, the Human Sciences Research Council, the South African Medical Research Council and the South African Labour and Development Research Unit. Data is available in the form of release surveys or aggregated data reports with varying levels of data quality.

The Education Management Information System within the Department for Basic Education in South Africa is responsible for developing and maintaining an integrated education database based on the collection of surveys of schools in the country. The types of data collected includes the School Master List, the Annual School Survey, Snap Survey as well as other data collected from various ad-hoc assessments conducted by the department (Department of Basic Education n.d.).

Statistics South Africa is responsible for the national collection of statistics in the country. The agency is working towards the construction of the National Statistical System which reports on data collected by the many departments within the country. The agency also provides various tools that allow one to interact with the Census, the Community Survey, household surveys and other sets of survey data. This information is published together with detailed accompanying metadata. The organisation's statistics cycle ensures that it produces new pieces of data on a weekly basis. Statistics South Africa has also put together the South Africa Statistical Quality Assessment Framework (SASQAF) which can be used to determine whether data produced by national departments or other institutes could be defined as official (Statistics South Africa 2008).

The Human Sciences Research Council is a statutory research agency and the largest dedicated research institute in social sciences and humanities in Africa. The research conducted by the Council is intended to inform public policy as well as monitor and evaluate policy implementation. The Council is responsible for various surveys that are conducted which follow the SASQAF guidelines of Statistics South Africa. Relevant surveys for the purposes of this study include the South African Social Attitudes Survey, the Trends in Mathematics and Science Study, the National Innovation Survey as well as the Health and Nutrition Examination Survey (Human Sciences Research Council n.d.).

The Southern Africa Labour and Development Research Unit based at the University of Cape Town has managed four waves of the National Income Dynamics Study (NIDS) which is a panel study of households in South Africa and studies the livelihoods and progression of their subjects over time. The themes covered in the various waves of the study track poverty, well-being, household structure, fertility and mortality, migration, labour market participation and economic activity, capital formation, health and education, vulnerability and social capital (Southern Africa Labour and Development Research Unit n.d.).

The South African Medical Research Council (MRC) was established to promote the delivery of health care and improve the public's quality of life in general by conducting research aimed at the key health challenges faced by the country. The MRC in partnership with the Human Sciences Research Council in 2002, 2008 and 2011 conducted the Youth Risk

Behaviour Survey (YRBS) across a selection of schools. The surveys focussed on risky behaviour, demographics as well as access to basic services in 2008 and 2011. In addition to the YRBS, in 2003, the MRC together with the Department of Health conducted the South African Demographic and Health Survey (DHS). For the latest iteration of the Demographic and Health Survey, the MRC has partnered with Statistics South Africa in order to roll out the 2015/16 version of the survey, the results of which are still pending (South African Medical Research Council n.d.).

5.5. Assess data quality using the PDQAF

In answering research question four of this study, the PDQAF is applied to test the data quality of all the selected data sources mentioned in the previous step. Using the elements, indicators and focal issue, a dataset can be assessed in how it fares from the perspective of a data user. Each focal point can be assessed to be positive, negative, neutral or is found to be not applicable. By uniformly applying this assessment to each focal issue within the Public Data Quality Assessment Framework, each dimension can thereafter be scored. A positive focal point carries a value of 1, a negative focal point carries a value of -1 and neutral/not applicable focal points carry a score of 0. As there are eleven dimensions within the framework, a survey could achieve between 11 and -11 points based on the number of positive or negative focal points per dimension. The scale is then rebased in terms of the number of applicable dimensions that affect the particular dataset. For example, if a survey has no data source system feeding the data collection effort, the traceability dimension is not applicable to the collection effort and is therefore excluded as a counted factor.

Each dataset has been assessed and the scoring of each dataset is provided in Appendices 2, 3, 4 and 5. These appendices provide a contextual assessment whereby each score is explained per focal issue depending on the characteristics of the data and the provided metadata. It is important to note, that the assessment is based on the information made available by the data providers to the public. Therefore, the data user’s perspective is crucial in making these determinations. In reality an organisation may follow a very developed data quality protection processes but if such information is not provided to the data user, the dataset will be assessed negatively with a finding stating that the metadata needs to be more thoroughly updated.

5.6. Data Presentation

Once the data is categorised, selected and sourced from the data provider, it is then possible to analyse the key trends found in the data. Many techniques are available for use which could provide valuable findings regarding the countries' functionings and capabilities. In order to exhibit the usefulness of the public data, a single functioning is selected. Following the identification of related datasets and the selection of data collected that was found to be of a higher level of data quality, datasets which ask similar questions and therefore are found to be comparable, are chosen for the purposes of this study. Care is made to present particular data trends together with the necessary supporting information to ensure that the reader is aware of the differences in the data collection strategy across the countries as well as to clearly identify the data source. Often the available data may also not represent a common time period. Therefore, such supporting details need to be also presented. Care is also taken to ensure that a common measurement scale is adopted to promote the comparison of the data amongst countries.

In the presentation of the data graphically, the following supporting details are included to clearly present the context of the information, viz., country, geographic granularity, time period, question asked to the respondent and the scale of measure for the particular responses.

5.7. Practical versus Conceptual Needs Assessment

Together with the presentation of the data, one must review the assessments of data quality and interpret the practical compromises against the conceptual requirements. Sen argues that the conceptual arguments should always trump the required practical compromises that emerge. In this regard, effort is made to ensure that the findings remain context-dependent and address the most pressing concerns which each country has raised. Furthermore, where data concerns are identified, these are documented and reported on. The findings address whether locally published data within Brazil, India and South Africa can be used to report on an indicator that best describes the school dropout phenomenon in these countries or if there are gaps pertaining to the valued functionings that are not supported by any data collection instrument. Where data gaps do emerge, these are documented and are included in the findings of this study as they are crucial in answering research question 4 which pertain to possible data concerns, which impedes data reporting.

In addition, each thematic area relevant to each functioning is assessed to determine if the specific questions or attributes collected within the chosen data collection instruments adequately express the depth and context of the specific functioning/theme/sub-theme. It is possible for a number of datasets to be identified in relation to a particular sub-theme without actually capturing the true meaning and context needed to be assessed. Therefore, to make such an assessment, the selected functioning is assessed to determine if the available data sources provide sufficient detail to describe all the facets of the related sub-theme.

5.8. Methodology Summary

The goal of the methodology is to merge data quality testing with Capability Approach theory. By including a practical assessment into the Capability Approach methodology, it is believed that the findings become more sensitive to the practicalities of data use in the identified three countries. It is this practical assessment which is lacking in operationalised Capability Approach implementations. In addition, some national data providers do not embrace the tenets of data quality in their data collection activities. It is believed that by exposing such limitations, focussed recommendations can be produced that identify for data providers, their particular data quality short comings.

To achieve this objective, the leading methods followed by Capability and Data Quality experts internationally have been adopted and applied specifically to Brazil, India and South Africa in the area of school dropouts. The steps outlined by Comim et al. have been followed by the HDCA in many of their capability related assessments, whilst the DQAF and SASQAF followed by the IMF and Statistics South Africa are the leading implementers of data quality assessments applicable to the selected 3 countries.

Whilst the theories selected have detailed academic foundations, the approaches adopted focus on the practical implementation, with the itemisation of specific steps. Therefore, the manner in which the frameworks are constructed allow for simple application and assessment. Furthermore, whilst the SDCAF is context dependent and informed by detailed literature in the school dropout conceptual space, the PDQAF is reusable in other public data quality assessments. Therefore, the PDQAF can be applied in any assessment of public data within the BRICS. Where a dimension is found to be inapplicable, the scoring system that has been formulated can be rebased to focus on only what is relevant.

Chapter 6

Construction of the School Dropout - Capability Assessment Framework

Following the outline of the Capability Set in section 3.4.3, this section details how the School Dropout Capability Assessment Framework (SDCAF) is constructed using the steps identified by Comim et al. (2008) specified within Chapter 5.

In answering research question two of the study pertaining to the development of a framework to assess the full range of factors that inform the learner's choice to drop out of school, the first task that must be followed involves assessing the state of disparity and need facing learners in Brazil, India and South Africa. A series of authors express their concerns, experiences and trends within the education sector from their countries. These are detailed within the Theoretical Framework in Chapter 2. These factors are assessed together with the outline of the SDCAF which is structured in section 3.4.3, are used to contextualise and explore each identified functioning in greater detail.

Once dimensions are fleshed out, the next task that is followed is the ranking of the identified dimensions in order of most pressing need. Whilst such a ranking of subjective concerns is difficult to objectively form, by using theories of human motivation such as Abraham Maslow's Hierarchy of Needs is useful in the ordering of importance of the school dropout functionings.

Following the finalisation of the SDCAF, appropriate datasets which describe the relevant sub-theme, theme and functioning combination are selected. Once datasets are identified, the relevant variables/questions from each dataset are explicitly identified. The data released by public data providers within each of the countries is assessed for relevance and fit. See Appendix 2 and 3 for the datasets that were identified and the questions per dataset which are relevant within this study.

6.1. Assessing learner's real freedom to attend school in preparation to access employment opportunities and improve their quality of life in Brazil, India and South Africa

As discussed in section 3.4.2, the central freedom which pertains to the learners of Brazil, India and South Africa involves exploring the extent to which learners attain a real freedom to attend school in a manner which adequately prepares them for employment opportunities and the ability to improve their quality of life. The extent to which this identified real freedom is attained is determined in the process of detailing the School Dropout Capability Set, which is instrumental to the SDCAF. The following sub-sections outline the relevant themes and sub-themes per functioning, as identified within school dropout literature discussed in Chapter 2.

6.1.1. Learners value being financially secure

Learners and their respective households value having a sense of financial security as discussed in sections 3.3.2 and 3.4.3. Poverty was found to be severely limiting factor which affects households forcing them to make trade-offs between the payment of school fees against purchasing food or supplies for the household. Furthermore, it was noted financial security enables the access to other valued functioning appreciated by learners which improve their general quality of life.

The financial security of the learners and their respective households have been identified as crucial in studies across the three countries. In general, school funding is based on the poverty level of the school's surrounding community. Whilst India and South Africa are concerned with funding the provision of education, Brazil has taken a step further in recognising the impact of financial means on attending school (see section 2.3). Due to this, access to state funded programmes to alleviate poverty are recognised as a key theme of financial security. Furthermore, the opportunity cost of learning compared to income received due to employment, is an important factor to include which was recognised by each country in sections 2.1, 2.2 and 2.3.

In South Africa (see section 2.1) general poverty is noted as a primary concern and this is also expressed in the family's inability to afford school fees, uniforms and various learner support materials that are required. The complicated situation facing child headed household was also noted in South Africa, where the child foregoes attending school to seek employment and generate income for survival purposes or to support younger family members.

In addition, the general socio-economic status of the learner, household and the community was identified as a key factor. This refers to the income of the household, their access to basic services and the type of dwelling the household resides in. Unemployed heads of households struggle to afford school fees or to provide for basic necessities such as school uniforms. Various authors also noted the importance of socio-economic conditions that the learner resides within in addition to the conditions faced at school by the learners.

The learner's household's access to basic services is also an important factor which can enable a learner's ability to study. Households without a suitable lighting source limit the learner's ability to study at home, additional factors such as access to a water source and suitable sanitation have a similar effect. Furthermore, the need for space in the household for studying is also an important factor.

Poverty is a complex subject affecting learners to different degrees. In schools in Brazil, India and South Africa programmes are offered to alleviate the difficulties associated with poverty, but not all learners have equitable access to these programmes. Often there is a stigma attached to poverty and the inability of a learner to afford such items required for school, resulting in poor learners' unwillingness to discuss with their schools their particular poverty status and therefore they do not access poverty alleviation programmes offered by a school.

In summary the key concerns that are raised are:

Financial Security	• Low household income • Inability to afford school fees and other resources required for school • Available space for home study and necessary resources • Socio-economic status of the learner, school and community • Cost of school fees compared to household income
State funded programmes	• Provision of state grants targeted at poor households • Provision of free schooling
Opportunity cost of school attendance	• Opportunity cost of school attendance in the form of employment income
Access to basic services	• Access to basic services

Table 22: Summation of issues of learners who value being financially secure

6.1.2. Learners value being physically well

As discussed in section 3.3.2 and 3.4.3, the value of bodily health was identified as one of the central human functional capabilities which includes issues of reproductive health, nourishment and shelter. The physical elements of well-being are distinguished from the mental elements as the factors are diverse and warrant separate analysis. When considering the centrality of this functioning in relation to human development, the specific impediments to school access attributable to poor health, must be explored further.

Malnutrition is closely linked to the issue of hunger and has a direct bearing on the learner's resultant dropout from school. Hunger and nutrition are essential concerns which universally affect poor communities as discussed in section 2.4. Factors such as the macronutrient and micronutrient intake of learners have an impact on school enrolment. Furthermore, as discussed in South Africa (see section 2.1), extreme poverty is also exhibited in occurrences of hunger and starvation. In Brazil (section 2.3) it was noted that children that have experienced hunger are more likely to leave school especially in urban areas. In India (section 2.2), it was found that lack of nourishment impedes academic performance therefore school nutrition or feeding programmes are particularly important, to alleviate the effects of hunger.

Internationally amongst medium income countries (see section 2.4), menstruation was recognised a crucial factor which is associated with high rates of absenteeism, which leads to school dropouts. This gender inequality is exacerbated by the costs of sanitary products, poor access to sanitation at school, a general lack of access to sanitary supplies and the absences of medication to treat menstruation discomfort as some communities view it as a 'luxury item'.

Teenage pregnancy is another extremely destabilising event in the life of female learners which has a direct bearing on their attendance of school across Brazil, India, South Africa and internationally (see sections 2.1, 2.2, 2.3 and 2.4). It was noted that pregnant learners often repeat grades and thereafter withdraw from school due to falling behind. This is especially common when the young female learner must also act as the primary caregiver for the child in her household. Furthermore, when the female learner's entry into school is delayed or they suffer grade repetition or a period of temporary withdrawal from school, such learners tend to be enrolled within the school system past puberty and into their late teenage years. The cumulative effect of these factors leads a female to be at a greater risk of falling pregnant. Some households withdraw teenage girls from school, to simply reduce the risk of pregnancy.

In addition, some learners engage in risky sexual behaviour as they are unaware of the dangers of such behaviour. Additional factors that were found linked to early pregnancy are a gender biased community, the household's financial strength, late school entry and a lack of access to contraceptives. Access to pregnancy prevention programmes which provide training to both male and female adolescents was also expected to curb rates of teenage pregnancy (see section 2.4).

In summary the key concerns that emerge are:

Malnutrition and Hunger	• Access to feeding programmes • Number of learners affected by malnutrition • Nutrition intake of learners • The effect of extreme poverty and hunger
Menstruation and Female Maturation	• Gender biased communities • Access to female sanitary products • Understanding female maturation • Cost of female sanitary products • Access to adequate sanitation • Provision of medication to counter menstruation discomforts
Teenage Pregnancy	• Provision of sex education • Pregnancy below the age of 20 • Accessibility and knowledge of contraception • Late school entry

Table 23: Summation of issues of learners who value being physically healthy

6.1.3. Learners value being mentally well

Capability Approach experts, as discussed in sections 3.3.2 and 3.4.3 noted that one's psychological well-being relates to feelings of peace of mind, lack of anxiety, happiness and feelings of harmony. The absence of such peace weakens a person's ability to function productively, therefore, the mental health of learners is a crucial element which is often not considered as a factor leading to learners dropping out of school, due to the intangibility of its impact.

Across Brazil, India and South Africa, the mental health of learners was found relevant (in sections 2.1, 2.2 and 2.3). Learners experienced feelings of shame when they were unable to pay for school fees or other resources required for school. This feeling of shame was found more apparent in schools with learners with varying degrees of income. Feelings of shame amongst learners was also experienced due to academic failure. In brief, the greater the experience of shame or stress felt from parental pressure, peer pressure, school pressure and external pressure add to a poor performing learner's loss of self-esteem and increases their anxiety level. Extreme circumstances lead to suicide.

An additional essential element related to being healthy relates to the effects of attention disorders, learning disorders and emotional disorders as discussed in section 2.4. A major barrier to inclusive education is found in the manner in which disabilities such as Attention

Deficit and Hyperactivity Disorder (ADHD) are diagnosed amongst the poorer communities. (see section 2.1). Mental ill health is also linked to poverty as the longer a person lives in poverty, the greater their risk of mental illness through exposure to stress, social exclusion, malnutrition, violence and trauma amongst other factors.

The recognition and diagnosis of a learning disability usually requires the assistance of a school psychologist. Poorer schools without the support of such services struggle to identify symptoms of the disorder. The disorder is commonly mistakenly identified as learning difficulty instead of as a learning disorder. Learners suffering from such disabilities are at a greater risk to dropout from school due to the greater and more frequent difficulties they experience (see section 2.4).

Emotional and behavioural disorders are an additional mental health factor which relates to school dropouts as discussed in section 2. 4. As similarly noted to the attention and learning disorders, they are not easily identified addressed amongst poorer schools. A learner's experience of a traumatic event or repeated exposure to highly stressful environment may lead to a psychiatric disorder. It was found that poorer communities require a greater integration of school staff and mental health professionals that can assist with the recognition and treatment of such.

In summary the key concerns that emerge are:

Psychological factors	• Social exclusion of poorer learners • Shame of poverty • Impatient teachers • Disinterested learners • Difficulty in learning complex subject matter • Teachers skills not sufficient for complex subject matter
Neurological factors	• Social exclusion of poorer learners • Shame of poverty • Impatient teachers • Disinterested learners • Difficulty in learning complex subject matter – teaching to memorise • Teachers skills not sufficient for complex subject matter • Support for Attention Disorders • Support for Learning Disabilities
Emotional factors	• Access to psychological services • Stress • Depression • Are there learning deficiencies experienced in reading, spelling, writing, comprehension, mathematics, problem-solving and attention • What are the experiences of trauma affecting the learner, i.e. crime, violence, abuse (domestic, family, sexual)

Table 24: Summation of issues of learners who value being mentally healthy

6.1.4. Learners value being taught in suitable school infrastructure facilities with necessary resources

From the Capability Approach literature discussed in sections 3.3.2 and 3.4.3 it was found that people value the facilities they work in and this was found to be a key driver of human development. When applied within the school context, the infrastructure such as the school buildings and facilities which the learners have access to, are enablers of building skills and knowledge, in this instance. Schools in a state of disrepair or without access to basic services severely inhibit the learner's learning experience.

This was found as significant in each of the countries (see sections 2.1, 2.2 and 2.3). Poorer schools struggle with the quality of infrastructure available to the learners, which includes factors such as having too few classrooms for the number of learners that are present, unequipped libraries and/or laboratories, lack of access to basic services such as water, electricity and sanitation and a lack of telecommunication devices.

It was also found that overcrowded classrooms limit the occurrence of effective learning. Furthermore, the authors noted how some schools mixed classrooms with learners of different ages, grade, abilities and subject area due to the limited space. Such environments tend to push out learners who are at risk when considered with other factors as well. Poor school infrastructure quality demotivates learners who are faced with challenges of poverty on a daily basis. The difficulties experiences relate to the unavailability of sanitation, water and electricity. In Brazil, it was found that the learner's desire to attend more dynamic and innovative schools based on their reputation was a unique theme to Brazil (see section 2.3).

In addition to lack of adequate sanitation facilities at a school, the lack of facilities and resources also reduce the likelihood that a learner will be suitably attracted to learning within the school environment. Schools with low security are generally poorly equipped due to effects of crime or the awareness of possible theft. Schools that lack of resources tended to suffered from higher rates of school dropouts. In addition, school facilities have a direct impact on the quality of education offered, therefore the availability and quality of supporting teaching and learning materials are important to foster an environment that is conducive to learning (see section 2.4).

In summary the key concerns that emerge are:

Improve Physical Infrastructure	• Ensure that the schools are of a suitable size with sufficient number of educators • School infrastructure maintenance • State of disrepair of the school
Effects of Physical Infrastructure	• School's access to basic services • Classrooms mixed by grade and subject area due to limited space • Reputation of the school to provide opportunities
Facilities and Resources	• Available facilities within a school • Security of the school • Available learning support materials • Available teaching aides • Provision secure environment

Table 25: Summation of issues of learners linked to infrastructure and facilities

6.1.5. Learners value being in a conducive home learning environment

In the Capability Approach literature, the need for an environment that is constructive and supportive fosters human development and enables a person to perform the tasks that they value (see sections 3.3.2, 3.4.3). When applied to the education sector, the learner's home is the enabling environment providing the learner the foundation to perform well in school. The absence of a supporting environment leads to constraints in their expression of personal convictions and generally limits their access to knowledge; nutrition and sound mental health.

As discussed in in Chapter 2, caregivers for learners inculcate a culture of learning if they value the educative process themselves. However, learners from poorer communities not only have to combat poverty and its immediate limitations, they often also contend with negative attitudes about learning and the value of school participation within the household or amongst family members. Such attitudes demotivate learners from effectively engaging with their school work. These factors are internationally recognised and common to Brazil, India and South Africa (see sections 2.1, 2.2, 2.3 and 2.4). Generally, learners are viewed as an asset to the household's progress and their education is considered to be empowering and a gateway to long term prosperous employment however, some households undervalue the benefits of education and pass this belief onto the learners. This lack of interest in education leads to absenteeism and dropping out becomes a common result.

The degree of importance attached to learning was closely linked to the level of education attained by the caregiver. Where the parent performed better in school, the learner received a higher level of support and consequently such households experienced lower levels of dropout. In South Africa, unemployed parents tended to pressure the learner (especially elder siblings) to leave school to instead seek employment that could supplement the household earnings (see section 2.1).

Other household factors include troubled parental relationships which are characterised by physical and verbal abuse directed towards the learner or amongst household members. Such an atmosphere is often considered as toxic for learning (see sections 2.1). The heightened levels of stress skew the learner's value system and motivation to continue school attendance. These factors are not only experienced at the household domain but are prevalent across communities, as also found in Brazil (see section 2.3).

A third key factor related to the household influences endured by the learner is the trauma from the deaths of parents or caregivers as well as the ordeals suffered when parents leave children unattended when they work far from home. Such learners carry the burden of becoming the head of the household and caregiver to younger siblings. These learners are more likely to dropout due to a lack of suitable role models who appreciate school attendance in the household. (see section 2.1). Therefore, in addition to the absence of an advising parent that can provide guidance with respect to school work, the learner is also forced to financially support themselves and other family members.

In summary the key concerns that emerge are:

Negative attitude to learning in the household, community	• Education attainment of the head of household and associated community • Head of household's opinion of education value • Personal and community's opinion of education value • Negative Environment
Abuse in the household and community	• Physical and verbal abuse • Substance (drug and alcohol) abuse within the household and community • Conduct disorders of learners
Child Headed Households	• Child headed households • Migrant caregivers • Household income of child headed households

Table 26: Summation of issues of learners who value a conducive learning environment

6.1.6. Learners value traveling to school in a safe and convenient manner

The ability to move freely and the ability to access the provided services was identified in the Capability Approach literature and discussed in sections 3.3.2 and 3.4.3. The freedom to move from place to place was found as a central human freedom, which is impeded due to the effects of crime which limits a person's ability to live free from harm. Amongst learner's this freedom is impeded to a greater extent amongst poorer learners who face difficulties when needing to travel great distances to reach the school.

Numerous authors in Brazil, India and South Africa, cite learners having to travel great distances to access school (see sections 2.1, 2.2 and 2.3). These learners have been found more vulnerable to school dropout. The cost, inconvenience and time that it requires to reach school are weighed against the perceived value of school and the immediate pressures such learners face. Internationally, it was expressed that the distance from school was an important school-related factor for learners leaving school. Traveling great distances is a factor which is compounded by higher transport costs which tend to affect poor households to a greater extent. Where learners are forced to walk great distances, they also tend to be tired before their school day begin, which has a demotivating effect on attendance (see section 2.4). Furthermore, in Brazil and India it was noted that young females are less likely to finish secondary schooling. The reasons identified in India related to a fear of rape when travelling to school. Caregivers consequently keep female learners at home to protect them from harassment (see sections 2.2 and 2.3).

In summary the key concerns that emerge are:

Distance to school	• Distance, time learner commutes to school • Available access road to the school • Urban/Rural location of the school
Provision of transportation	• Available transportation services • Cost of transportation
Sexual harassment of female learners	• Female learners suffer sexual harassment in commute to school

Table 27: Summation of issues of learners who value safe and convenient travel to school

6.1.7. Learners value meaningful participation in school

The provision of education and participation in school is recognised as a central enabler of capabilities of a person to achieve their valued aspirations in life. Furthermore, social opportunities such as education and health enable a person to be productive participants in the economy at a later stage in their lives. Therefore, education is viewed as a gateway to achieving a valued outcome and is not an outcome on its own as discussed in sections 3.3.2 and 3.4.3. An evaluation of one’s educational aspirations should be in terms of the learning outcomes that one would be able to achieve, as with greater participation it will be important to determine what one’s real educational choices are (see section 3.3.1).

Across Brazil, India and South Africa, the learner’s ability to contribute and participate productively during a lesson is an important determinant of how a learner values their attendance of school Learners that struggle with understanding elementary level content are more likely to leave school before the tenth grade. These trends continue to persist when learners are promoted and have not attained the necessary understanding of the content of the particular grade. In Brazil, low academic achievement in school had a direct bearing on their inclination to leave school (see sections 2.1, 2.2 and 2.3).

The overall quality of the education system is closely bound to the quality of the country's teachers. However, defining what constitutes teacher quality is difficult to express. The best set of factors that were defined are teacher professionalism, having a desire to teach, the pedagogical skills and knowledge for teaching and the ability to impart and instil knowledge, skills and values to the learners. (see section 2.1).

An additional factor affecting the efficacy of a learner's participation in school is their level of interest in learning. The learner's level of interest in learning depends closely on the pressures the learner faces from their household or community environments. Crime, violence, substance abuse and other negative environmental factors can impede a learner's ambition to participate in school (see sections 2.1, 2.2 and 2.3).

The final major concern influencing the learner's effective participation in school is the issue of discrimination. Those that were identified as being affected by such issues are pregnant learners, marginalised learners in wealthy schools and minorities. In India, it was also noted that Muslims (both male and female) and female learners are disenfranchised as they were found to be in the minority as discussed in section 2.2. In India, gender sensitivity training in schools was suggested to promote gender equality.

In summary the key concerns that emerge are:

Comparative academic strength of the learner	• Understanding elementary level concepts • Comprehension of classwork
Teacher Quality	• Teacher professionalism • Pedagogical skills, training and knowledge for teaching • The ability to impart and instil knowledge, skills and values to the learners
Teacher Motivation	• Lack of teacher motivation • Positive and negative teacher incentives
Household and Community Environment	• Learner's motivation for school work • Crime, violence, substance abuse and other negative environmental factors • Socio-economic conditions

Discrimination against some learners	• Discrimination in the household, school and community • Report school attendance trends by: ○ Gender ○ Race ○ Caste ○ Creed • Sensitivity training pertaining to marginalised groups in the curriculum

Table 28: Summation of issues of learners who value meaningful participation in school

6.1.8. Learners value being able to freely express their opinions

The ability to freely express an opinion is identified in the Capability Approach as such a freedom is a valued state all people wish to attain. It is linked to a person’s civil rights. The lack of censorship and political freedom is part of our inherent need for social participation. Restraints on one’s expression limits open discussion and debate. (see sections 3.3.2, 3.4.3).

There are various factors which limit the learner’s ability to express themselves such as the effect that a negative attitude of an educator which curtails learner involvement in a lesson. Such an attitude influences the learner’s interest in school attendance which leads to their dropout. In addition, sometimes teachers develop a preconceived attitude to a particular learner, which intentionally or otherwise discriminates that particular learner despite the value of their contribution to the lesson. On the learner’s side, the lack of discipline amongst learners impacts the educator’s morale, leading to a negative atmosphere which further diminishes the quality of the lessons (see section 2.1).

Some school authorities tend to restrict free speech despite constitutional protections with the aim to maintain discipline and orderliness in a school. Pedagogical research highlights that it is important for the learner to freely express themselves as it fosters their cognitive development and their appreciation for school enrolment. In addition, school uniform policy in schools require that learners comply with rules which may disavow their wanting and right to use religious attire. Such restrictions leave learners feeling attacked and therefore more likely to leave school (see sections 2.1, 2.2 and 2.4).

In summary the key concerns that emerge are:

Negative teacher attitude	• Restrictive classroom environment • Teachers with preconceived attitudes to learners
Conducive school learning environment	• Undisciplined learners • Appropriate number of teachers
Restricted cultural expression	• Restrictions on cultural attire • Promotion of Cultural Diversity

	• Restrictions on free speech

Table 29: Summation of issues of learners who value free expression of their opinions

6.2. Organise the Dimensions in order of the most pressing need

Following the discussion of the previous section, the outlined dimensions are (1) learners value being financially secure, (2) learners value being physically well, (3) learners value being mentally well, (4) learners value being taught in suitable school infrastructure facilities with necessary resources, (5) learners value being in a conducive home learning environment, (6) learners value traveling to school in a safe and convenient manner, (7) learners value meaningful participation in school and (8) learners value being able to freely express their opinions. Comim et al. (2008) recommends that the dimensions that are selected should be ranked to identify the order of importance in terms of what learners would find most valuable. However, this ranking is primarily done in order to better arrange survey instruments for data collection (see task 3 in section 5.2). In this particular study, we will not produce primary data but are rather evaluating secondary data when applying the PDQAF in Chapter 7. However, in order to evaluate the relevance of the available data it will be important to recognise in what order these dimensions can be positioned to better understand the importance of the identified dataset.

If we apply Abraham Maslow's hierarchy of needs, we can begin to place the dimensions into order following Maslow's identified five hierarchical categories of needs viz., (1) Physiological Needs, (2) Safety and Security, (3) Love and Belonging, (4) Self-esteem and (5) Self-actualisation. The first four categories describe deficiencies that a person may face. Only when these deficiencies are catered for can one attain self-actualisation. Each category, enables the following the category, therefore the physiological need (the need to breathe, eat, sleep, etc.) trumps the need for safety and security (health, employment, stability, etc.). It is this concept of 'pre-potency' that Maslow introduces which allows one to rank categories that are considered subjective and incomparable (Griffin 1998).

In light of Maslow's hierarchy, the foundational functioning, upon which the others with the Capability Set stand is learner's value of physical well-being. The learner's health requirements are consistent with Maslow's statements on physiological needs and also pertains to feelings of safety and security. Learners faced with starvation and malnutrition can become consumed by this need, as are the learners that face issues of female maturation and menstruation. The combined effect of such factors warrants the placement at position 1.

The learner's financial security is the next most important dimension. Available income and costs of school fees and other resources impact one's ability to access food, water, shelter, etc. Therefore, the learner's household's ability to attain employment and to earn money is placed in the second place within the Capability Set, as one's ability to care for themselves is linked to concerns about one's safety and security.

The mental well-being of the learner has been placed in the third highest rank. The psychological, neurological and emotional factors that are linked to mental health concerns are closely linked to issues of health and social stability. Such stresses impede one's physical health, and are also related to one's sense of inclusion within their household, classroom and community which relates both to Maslow's second (Safety and Security) and third tiers (Love and Belonging).

The fourth priority for learners pertains to their need for safe and convenient travel to school. As learners often traverse long distance becoming quite tired and hungry whilst facing various perils on their journey to school, we find that their security is jeopardised (a safety and security issue) as well. Depending on the levels of hunger and tiredness arising from the learner's travel to school, one could make an argument for the requirement to be placed as the third priority due to the greater health significance of the dimension. However due to the difficulty to measure the impact on health requirements, it is decided to place this as a fourth priority.

The need for a conducive home learning has been placed as the fifth most important functioning within the Capability Set due to the impact it makes on the learner's family life, and their general social stability within the home which corresponds with Maslow's concept of safety and security. The family element relates to the Love and Belonging need that Maslow identified. The connection that is built with members of the household and their influence on the learner tends to emphasise Maslow's third tier.

The impact of the quality of the school infrastructure and supporting resources identified as the next most important functioning. Whilst the school provides the learner with a sense of safety which is an early foundational need identified by Maslow (Safety and Security), the learner's appreciation of the facilities relates to the learner's perception of experience of attending school and learning which is linked to Maslow's final tier of Self Actualisation. This counter-balancing of safety concerns and health supported by the experience of school attendance place the value of school infrastructure as only the sixth most important concern.

The learner's ability to freely express themselves is related to their connected sense of belonging to the school, their confidence and need to be unique which relates to Maslow's third tier on Love and Belonging and the fourth tier on Self Esteem which place the learner's free expression in the position of rank 7.

The final need which may be found to be surprisingly last in importance is the value the learner derives from participating in class. The acceptance of being in school and attaining meaning from the experience to fulfil one's inner potential is linked to the Maslow's final need of Self-Actualisation. Achieving self-actualisation according to Maslow is only possible when all lower tier dimensions are supported.

From this application of Maslow's hierarchy, one can determine that the value the learner place on school participation is far outweighed by their needs for health, shelter, safety and free expression. The learner can only begin to value school participation once these preceding needs are accommodated. Each of the functionings identified have individual thematic concerns, which can each be ranked following Maslow's approach, however the application at this level enables a rudimentary ordering of functionings which will be useful when assessing the value of the data available per dimension. A more granular ranking is not required at this stage.

6.3. Outline and Order of the School Dropout Capability Assessment Framework

Using the tenets of the Capability Approach, one is able to express the attainment of the key real freedom related to the School Dropout Phenomenon in terms of a Capability Set. Therefore, the learner's real freedom to attend school, which readies the learner for employment and a possible future comfortable life can be described in terms of functionings which describe the learners' current valued states of being and doing. These functionings can be described in terms of Clark's characteristics of dimensions. Clark noted that functionings relate to the survivalist development requirements which tend to affect disadvantaged communities to a greater extent, factors for mental well-being and one's enjoyment or appreciation of their activities. Each of these factors were examined and a selection of functionings most appropriate were made together with the identification of key themes which apply to the school dropout phenomenon. The functionings that emerged in relation to the survivalist perspective were functionings of being financially well, being physically well, being taught in suitable infrastructure with necessary resources and the learners ability to travel to school safely and conveniently. The functionings related to mental well-being followed being mentally well considering psychological, neurological and emotional factors as well the state of the learning culture and environment the learner found themselves in. The enjoyment or appreciation perspective related to the learner's ability to meaningfully participate in school and their ability to freely express themselves.

Each of these functionings are found applicable within the literature describing school dropout factors in Brazil, India and South Africa. However, by applying the Capability Approach, to the literature describing school dropouts, one is able to describe each of these functionings in terms of what learners and their respective communities value most. An interesting finding is that the experiences across Brazil, India and South Africa differ and therefore certain factors are not universally recognised in literature in Brazil, India and South Africa despite their pertinence. For example, where the development challenges are stark, the tangible concerns are discussed in detail in the literature. However, intangible factors such as mental well-being are not universally noted.

Through the application of Maslow's hierarchy of needs, each functioning is crudely ranked presenting an order of the importance which learner's attach to such themes. Using Maslow's hierarchy it is clear that physical well-being becomes a more crucial factor to the learner than achieving meaningful participation within class. The ranked collection of functionings, detailed in Table 30, is used to identify relevant datasets within Brazil, India and South Africa. The next chapter outlines the data quality assessment factors which should be applied to test whether such datasets can be used for reporting. Within Chapter 8, datasets are identified for a selection of these functionings and themes and the data quality assessment framework is applied to each to determine if the identified data is of suitable quality for reporting on school dropout factors.

Capability: Learner's real freedom to attend school in preparation to access employment opportunities and attain comfortable living

Functioning	Theme	Sub-Theme
Being physically well	Malnutrition and Hunger	Access to feeding programmes
		Number of learners affected by malnutrition
		Nutrition intake of learners
		The effect of extreme poverty and hunger

Functioning	Theme	Sub-Theme
	Menstruation and Female Maturation	Gender biased communities
		Access to female sanitary products
		Understanding female maturation
		Cost of female sanitary products
		Access to adequate sanitation
		Provision of medication to counter menstruation discomforts
	Teenage Pregnancy	Provision of sex education
		Pregnancy below the age of 20
		Accessibility and knowledge of contraception
		Late school entry
Being financially secure	Financial Security	Low household income
		Inability to afford school fees and other resources required for school
		Available space for home study and necessary resources
		Socio-economic status of the learner, school and community
		Cost of school fees compared to household income
	State funded programmes	Provision of state grants targeted at poor households
		Provision of free schooling
	Opportunity cost of school attendance	Opportunity cost of school attendance in the form of employment income
	Access to basic services	Access to basic services
Being Mentally Well	Psychological factors	Social exclusion of poorer learners
		Shame of poverty
		Impatient teachers
		Disinterested learners
		Difficulty in learning complex subject matter
		Teachers skills not sufficient for complex subject matter
	Neurological factors	Social exclusion of poorer learners
		Shame of poverty
		Impatient teachers
		Disinterested learners
		Difficulty in learning complex subject matter – teaching to memorise
		Teachers skills not sufficient for complex subject matter
		Support for attention disorder symptoms
		Support for learning disability symptoms
	Emotional factors	Access to psychological services
		Depression
		Stress
		Are there learning deficiencies experienced in reading, spelling, writing, comprehension, mathematics, problem-solving and attention

Functioning	Theme	Sub-Theme
		What are the experiences of trauma affecting the learner, i.e. crime, violence, abuse (domestic, family, sexual)
Traveling to school safely and conveniently	Distance to school	Distance, time learner commutes to school
		Available access road to school
		Urban/Rural location of the school
	Provision of transportation	Available transportation services
		Cost of transportation
	Sexual harassment of female learners	Female learners suffer sexual harassment in commute to school
Being in a conducive home learning environment	Negative attitude to learning in the household, community	Education attainment of the head of household and associated community
		Negative environment
		Head of household's opinion of education value
		Personal and community's opinion of education value
	Abuse in the household and community	Physical and verbal abuse
		Substance (drug and alcohol) abuse within the household and community
		Conduct disorders of learners
	Child Headed Households	Child headed households
		Migrant caregivers
		Household income of child headed households
Being Taught in Suitable Infrastructure with necessary resources	Improve Physical Infrastructure	Ensure that the schools are of a suitable size with sufficient number of educators
		School infrastructure maintenance
		State of disrepair of the school
	Effects of Physical Infrastructure	School's access to basic services
		Classrooms mixed by grade and subject area due to limited space
		Reputation of the school to provide opportunities
	Facilities and Resources	Available facilities within a school
		Security of the school
		Available learning support materials
		Available teaching aides
		Provision secure environment
Free Expression of Opinions	Negative teacher attitude	Restrictive classroom environment
		Teachers with preconceived attitudes to learners
	Conducive school learning environment	Undisciplined learners
		Appropriate number of teachers
	Restricted cultural expression	Restrictions on cultural attire
		Promotion of Cultural Diversity
		Restrictions on free speech
		Understanding elementary level concepts

Functioning	Theme	Sub-Theme
Meaningful participation in school	Comparative academic strength of the learner	Comprehension of classwork
	Teacher Quality	Teacher professionalism
		Pedagogical skills, training and knowledge for teaching
		The ability to impart and instil knowledge, skills and values to the learners
	Teacher Motivation	Lack of teacher motivation
		Positive and negative teacher incentives
	Household and Community Environment	Learner’s motivation for school work
		Crime, violence, substance abuse and other negative environmental factors
		Socio-economic conditions
	Discrimination against some learners	Discrimination in the household, school and community
		Report school attendance trends by Gender, Race, Caste, Creed
		Include sensitivity training pertaining to marginalised groups in the curriculum

Table 30: School Dropout Capability Assessment Framework

Chapter 7

Construction of the Public Data Quality Assessment Framework

The second phase of the assessment in answering research question two of the study in terms of the necessary frameworks to assess school dropout concerns, requires the development of the PDQAF. This framework assists public data users from Brazil, India and South Africa to assess the data quality of publicly available data. As discussed in Chapter 4, the outline of the PDQAF is based on the Sadiq-Eppler data quality dimensions, the IMF DQAF and Statistics South Africa's SASQAF. Table 18 in Chapter 4 details the selection of dimensions which forms the basis of the framework. Each dimension can thereafter be decomposed into a selection of elements, indicators and focal issues. Each focal issue is then described in terms of positive or negative traits which assist in determining whether or not high quality data is provided.

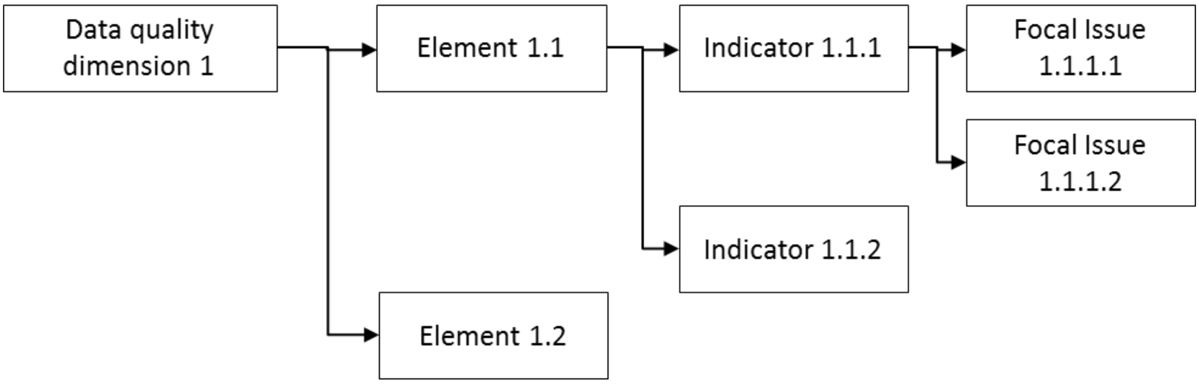


Figure 5: Iterative process for defining data quality dimension (Adaptation of IMF DQAF and SASQAF)

7.1. Organisational Dimensions of Data Quality

Sadiq describes the first perspective that is considered when reviewing data quality is the organisation level wherein management strategies are put into practice, definitions are decided upon and the use of the data is determined. Each of these decisions impacts the operational delivery of the data and must be concisely captured within the metadata for every dataset. Not only is metadata useful to the organisational employees, it is also crucial to the data user as it is the solitary form of communication between the data provider and the data user regarding the data collection and quality protection processes that were instituted.

7.1.1. Dimension 1: Metadata

When one assesses the core concerns that emerge from the organisational perspective, one finds the role of providing concise and comprehensive meta data that can be interrogated by the data user as crucial to the data user. Whilst the data provider may attempt the best possible data management approach to ensure he principles of data quality are upheld, from the perspective of the data user, the attained level of data quality can only be critiqued by the contents of information accessible to them. Therefore, the metadata is not only extra documentation which accompanies a dataset but, in effect, is a practical guide for the data user to peruse and to form an opinion regarding the dataset's inherent data quality.

As discussed in section 4.2, the metadata must capture three crucial elements (1) describe the content of what is included within the published dataset, (2) communicate the context from which the dataset arises and (3) discusses the technical structure of the dataset. These three points provide the context for the necessary elements of this dimension.

Element 1: Does the metadata convey the contents of the dataset?

In determining whether the metadata has suitably described the contents of the dataset, various authors offer the following points that the data user values when examining the metadata. Firstly, the metadata must provide definitions pertaining to concepts and classifications included within the metadata documentation (see section 4.5.1). Where there are deviations from the accepted norm for a particular definition, such occurrences should be noted within the documentation. Secondly, metadata needs to be kept up to date and in line with the latest revisions related to a particular data collection. Where changes are made to a dataset, these revisions need to be traced. Thirdly, the data user must be informed about the rationale for each table and each field that is included within the tables. Providing such details allows the data user to identify which sets of information is useful for their purposes and assists with general data perusal. Lastly, it is very important for the data user to understand the geographic granularity of the data that is provided. The spatial context of data is extremely useful when evaluating public policy (see section 4.2). These points are structured in the following table.

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are concepts and definitions and classification provided within the metadata to describe the underlying data?	• Are definitions made available within the metadata? • Are deviations from standards reported?	• Concepts, definitions, classifications and examples are provided. • All concepts used are well defined and	• Few or no definitions are provided • No administrative records are kept or they are poorly maintained

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
		documented in administrative records	
Is the metadata up to date?	• Is a document register regularly maintained in line with the data collection strategy?	• Documentation is regularly updated	• Documentation is infrequently or not updated.
Are all tables and fields defined?	• Is the purpose and definition of each table and field within the dataset reported on?	• Each table and field is well defined • Differences between similar fields and/or tables are described	• Limited effort is made to provide such details to the data user
Is the geographic distribution clearly defined?	• Is the geographic level of data granularity described within the metadata?	• The granularity is documented • Changes to the granularity or geographic boundaries is referred to within the documentation	• Such details are not provided or provided in an incomplete fashion

Table 31: Indicators describing whether the metadata suitably conveys the contents of the dataset

Element 2: Does the metadata describe the context of the dataset?

As discussed in sections 4.2 and 4.5.1, the metadata should also be made up of administrative information which describes the context within which the data is produced. Such metadata should include information regarding how and when the dataset was produced, how it is managed and the technical details about the methodology that was followed when producing the dataset. The metadata should also discuss the methods and statistical techniques used in producing the particular published dataset. In communicating how the data quality is managed within an organisation, the key tasks for ensuring data quality should be documented. These tasks include defining how the data quality is defined, how data quality is measured by the data provider, the processes followed in analysing and processing data quality concerns and how the organisation controlled data quality during their management tasks.

In addition to describing the data quality management processes followed by the organisation, the metadata must also suitably convey the methodology followed by the organisation during data collection. Both the techniques employed as well as the sources that are used to collect data should be documented. Furthermore, the metadata should capture the norms and standards related to such data collection activities. A key concern that must be addressed and captured within the metadata (in relation to surveys carried out) is the sampling selection and questionnaire design. The metadata should communicate to the data user how such tasks were carried out, to ensure that such tasks are methodologically sound (see section 4.5.4). Often public data is collected in a survey and an effort must be made to anonymise the data to protect the identities of the survey respondents The manner in which this is carried out also requires documentation. Furthermore, the scope of the particular dataset needs to be outlined, thereby informing the public of the how representative the dataset is.

In addition to providing raw data, in order for the data user to understand the context of the data's use, it is useful to accompany the dataset with a report on the particular findings related to the data collection. The findings related to the dataset help the organisation understand how to improve future data collections and present a preliminary data position, which data users can attempt to instantiate or disprove (see section 4.5.1).

Lastly, the metadata must provide details pertaining to the context of the dataset and ensure that this information is understandable to a lay-person, balancing the needs of the technical against how understandable the information becomes (see section 4.2).

Indicator	**Focal Issue**	**High Quality Trait**	**Poor Quality Trait**
Does the metadata describe the data quality practices applied when producing the dataset?	• Does the organisation define data quality within the metadata? • How does the organisation measure data quality within the metadata? • How does the organisation process data concerns? Is this discussed in the metadata? • Are data quality controls discussed within the metadata?	• Provides adequate or partial documentation on each of the focal issues discussed • Provides information on best practices and scope of data quality management activities	• Limited or no documentation on the focal issues are provided. • Limited information provided on what the best practices are pertaining to data quality management.
Does the metadata comprehensively describe the data collection methodology?	• Are the statistical techniques employed suitably documented? • Are all data sources used documented? • Is the sampling selection framework and decisions taken adequately documented? • Are the techniques for ensuring anonymity provided? • Is the questionnaire design documented? • Are the norms and standards regarding the data collection discussed within the metadata? • Is the scope of the data collection documented?	• Statistical techniques are discussed in adequate detail • Data sources are referenced • The sampling framework is comprehensively provided ensuring that the survey is representative of the geographic distribution • The anonymization techniques are well documented • The questionnaire design and layout is adequately reported on • The norms in relation to the dataset subject	• Statistical techniques are weakly reported on • Data sources are not completed referenced • The sampling framework is not adequately provided • The anonymization techniques are incomplete or excluded • The questionnaire design is excluded or does not cover all aspects of the survey • The norms in relation to the dataset subject matter are not documented, are outdated, are incorrect or excluded

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
		matter are documented • The scope of the dataset is clearly delineated	• The scope of the dataset is not specified
Is there an accompanying findings report?	• Is a data output report provided together with the dataset publication?	• A findings report is made available and provides a detailed analysis of the uses of dataset and its limitations	• A findings report is not produced or is limited in its analysis
Is the metadata clear and understandable to the data user?	• Does the metadata concisely and comprehensively describe how the dataset is produced?	• The language used is not heavily dependent on internal jargon • The documentation is provided in a logical manner • The documentation is well referenced • The documentation is current	• The language used is not understandable to the lay-person • Information is presented in an unstructured format • Limited or no references are made to data • The documentation is outdated and irrelevant

Table 32: Indicators describing whether the metadata suitably conveys the context of the dataset

Element 3: Does the metadata explain the structure of the dataset?

The third form of metadata refers to the structural elements of the dataset. This metadata assists the data user in understanding each individual sub-component represented within the dataset and how these components interrelate and are combined into the published dataset (see section 4.2). Such information is generally technical in nature but provides the data user the specifics of the structure of each table and field and explains the logical rules inferred by the manner the data is structured. This involves providing details about the data types that are involved, the digital size of the information or in the cases of published databases, the metadata must detail how to physically access the information stored within the available data tool. The metadata will also need to detail the hardware and software requirements for accessing the particular data tool and rendering a report.

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Does the metadata document the structure of the dataset?	• Is the physical layout (data tables, fields, database) of the data structures documented? • Are the hardware and software requirements	• The physical layout is provided with details of data types, structures • Hardware and software requirements are documented	• The physical layout is not provided with sufficient detail or are outdated or inaccurate • Hardware and software requirements are

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
	detailed regarding use of the data? • Does the metadata explain how to navigate and use online databases if relevant to the data publication	• A user guide is provided explaining to the data user how to access the relevant data	excluded or are inaccurate • A user guide is inaccurate, outdated or non-existent

Table 33: Indicators describing whether the metadata suitably conveys the structure of the dataset

7.2. Architectural Dimensions of Data Quality

The architectural perspective is related to how the organisation puts into practice the data quality principles that are identified by management. These principles relate to the community level, the product level and the infrastructure level of data quality. Where these principles affect the data user, they have been selected for use in the framework.

7.2.1. Dimension 2: Comprehensiveness

As discussed in section 4.3, the principle of comprehensiveness relates to completeness and the sense of surety that the data user finds when working with the relevant components and values are present within the dataset in line with the scope of the particular data collection. Comprehensiveness in this sense refers to the entire data publication and not a single field or subsection within the data publication. When discussing Comprehensiveness, two key elements emerge within the literature, firstly whether the data physically possesses all the data values and components that are outlined as part of the reported data structure and secondly, whether the business rules have been met which outline what the dataset should contain.

Element 1: Does the dataset contain all required statistical units?

When assessing the coverage of a dataset, one should determine if all the statistical units that should be covered are included. For example, if school enrolment data is reviewed, one should determine if the geographic granularity is represented and whether dimensions of gender and grade are populated. In these instances, the metadata will assist in outlining what fields should be included and populated within the dataset. Therefore, when reviewing the metadata of the published data, a comparison must be made to assess whether the structure provided in the dataset corresponds with the reported structure. Furthermore, when assessing completeness, one should determine if the available data provides a reasonable approximation of data based on the definitions, scope, classification and timelines of the data publication (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are all expected statistical units populated?	• Is each combination of statistical units represented within the data as per the required scope of the dataset?	• All data combinations are present in the dataset in line with all proviso's stipulated within the metadata	• Not all data combinations are present in the dataset
Does the dataset structure match the metadata outline?	• Does the dataset layout match the metadata in terms of the table, field names provided?	• All fields and tables are consistent with the metadata structures	• The fields and tables provided in the dataset are inconsistent to the metadata or are not present
Does the dataset reasonably convey the definitions, scope, classification, valuation and timeline as specified within the metadata?	• Do values provided within a field broadly match the definition? • Are the provided Statistical units matching the scope of the dataset? • Do the categories within the data match the specified classification? • Does the data match the data type of the particular fields? • Does the time periods within the data match the expected time period?	• The dataset agrees with each of the issues raised within the focal points	• The dataset disagrees with the issues raised within the focal points

Table 34: Indicators specifying whether the dataset contains all required statistical units

Element 2: Are all data rules met within the dataset?

It was noted that a dataset is 100% complete if every business rule is met. Therefore, to measure completeness, one would need to quantify the number of blank values where there were expected entries of information (see section 4.3). In addition, one should assess if the mandatory fields and inapplicable fields are updated in an appropriate manner. Mandatory fields should be populated, whilst inapplicable fields such as the school name of learners not attending school for example should be blank. The degree to which these basic constraints are upheld will assist in determining if the dataset is complete. It should be noted that when assessing completeness, the intention is to only determine if the necessary values are populated but not to determine if they are correct.

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are there blank data entries where the data rule mandates an entry?	• As per the data rules, are all necessary fields provided and populated?	• Each record has populated entry in keeping with the data rules	• Not every record has a populated entry in keeping with the data rules
Are all mandatory attributes populated?	• Are Mandatory fields populated?	• There are no blank entries where there are expected values	• There are blank entries where there are expected values
Are all inapplicable attributes left blank?	• Are inapplicable attributes blank where appropriate	• Appropriate entries are blank as per business rules	• Expected blank entries are incorrectly populated

Table 35: Indicators specifying whether all data rules are met within the dataset

7.2.2. Dimension 3: Accuracy

As stated within the literature review, accuracy has been defined as a state the data achieves when it is free from defects. Accuracy can be assessed when one compares the difference between the expected value and the provided value. The relevant elements of data accuracy relate to (1) data coverage and (2) whether the data is generally suitable for reporting (see section 4.5.4).

Element 1: Are the data coverage methods adequate?

When assessing the data coverage of the dataset, where data sampling is conducted, one should assess the data sampling errors that were calculated. Such information needs to be provided within the metadata of the dataset. These include assessing the standard error, the coefficient of variation, the confidence interval and the mean square error. These calculations need to fall within the international norm which should also be referred to in the metadata. Where data sampling is not utilised, one needs to assess how well the target population is represented in the particular data collection, as well as to identify which imputation techniques were used and if they were adequately used. Furthermore, non-sampling error measures need to be assessed in terms of the international standards. These particular error calculations include frame coverage, duplication in the frame, number of statistical units out of scope misclassification errors, measurement errors, processing errors and model assumption errors (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Has the dataset adequately applied data	• Are the standard error, the coefficient of variation, the	• Data sampling calculations are	• Data sampling calculations are

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
sampling techniques?	confidence interval and the mean square error calculation in line with international norms and standards	within international norms and standards	within international norms and standards
Are imputation techniques adequately applied	• Is the response rate too low and is imputation rate too high?	• The response rate is equal or above the international standard	• The imputation rate is greater than the international standard
Has the dataset adequately applied non sampling techniques	• Is the frame coverage, duplication in the frame, number of statistical units out of scope, misclassification errors, measurement errors, processing errors and model assumption errors data calculations in line with international norms and standards?	• Non sampling calculations are within international norms and standards	• Non sampling calculations are within international norms and standards

Table 36: Indicators specifying whether data coverage has been adequately met

Element 2: Is the data suitable for reporting?

To determine the suitability of data for reporting, the dataset must be assessed in terms of how the data agrees with a comparative data source when compared at a particular granularity. When a dataset is dependent on another data source, this data source must also be vetted in terms of the PDQAF criteria. Lastly, a key task that must be completed is to assess whether the data maintenance procedures in place are adequate. These relate to how the sample frame is updated, whether the quality assurance processes are maintained and whether a data audit provides a clean report in terms of the types of errors produced internally (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Does the provided data correspond with a comparative source?	• The aggregated data of the dataset equates to the comparative data source	• The difference between data sources is within acceptable standards	• The difference between data sources are beyond acceptable standards
Has a downstream	• Was the primary data source assessed in line with the	• The primary source was assessed and	• The primary source was not assessed or

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
source been quality assessed?	Public Data Quality Assessment Framework	found to be a quality product	was not found to be a quality product
Are the adopted maintenance procedures suitable?	• Has the sample frame been regularly updated and managed? • Does the organisation conduct regularly quality assurance procedures? • Does the organisation conduct regular data audits and are the errors identified acceptable?	• The sample frames are thoroughly documented and regularly maintained • The organisation has adopted quality assurance procedures • The organisation subscribes to data audits and the errors produced are within acceptable international standards	• The sample frames are not adequately documented and are not regularly maintained • The organisation has not adopted quality assurance procedures • The organisation does not subscribe to data audits or the errors produced are beyond acceptable international standards

Table 37: Indicators specifying whether the data source is suitable for reporting

7.2.3. Dimension 4: Clarity

Clarity refers to how well a term is used within a dataset and understood by the data users. The degree of how well such terms are understood by the public drives communication between the data provider and the data user. Well clarified definitions are a product of a well-functioning naming convention system within the organisation. One can measure clarity if one compares the number of terms used within the dataset that follow the organisation's defined naming conventions against the number of terms used that do not follow such principles (see section 4.3). A clear and understandable dataset is one that is released together with metadata that includes information pertaining to the scope, statistical techniques employed, the standards the dataset should be assessed in comparison to and the publication of statistical findings based on the released data (see section 4.5.4). Note that these particular concerns are not included as part of the clarity dimension as they are assessed in the metadata dimension.

Element 1: Is there consistency between terms used and the naming convention?

The dimension can be measured in terms of how aligned the terms used in a dataset are to the naming conventions of the organisation. The chosen approach in measuring the dimension is to determine whether the data provider publishes information on the concepts, definitions, classification and the standards that apply to the particular definition (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Is there consistency between the terms used and the organisation's definitions?	• Are concepts, definitions, classifications and standards applicable to the dataset made available? • Are the terms used based on agreed company definitions?	• Concepts, definitions, classifications and standards are made available to the public that are relevant to the dataset • Terms are used that agree with company definitions	• Concepts, definitions, classifications and standards are not made available to the public that are relevant to the dataset • Terms are used which are inconsistent with company definitions

Table 38: Indicators specifying whether the data source consistently uses terms in line with organisation naming conventions.

7.2.4. Dimension 5: Applicability

As described within the literature, the applicability of the dataset must be examined in terms of how useful it is found to be by data users. The contents of the dataset must be streamlined to ensure that all superfluous components which are not beneficial to the data user should be dispensed with (see section 4.3). In addition, the dataset in its entirety must be analysed in terms of why it was produced and whether it meets the needs of the data user and is therefore beneficial to them (see section 4.5.4).

Element 1: Does the dataset meet the needs of the data users?

To assess applicability, the data users for the particular dataset need to be identified and their reasons for accessing the dataset must be determined. Furthermore, the user's requirements for such data needs to be assessed and it must be determined if the dataset provided caters to these requirements (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Does the dataset meet the requirements of the data users?	• Have the public's data requirements been determined and have they been made available? • Do the identified data requirements correspond with the purpose of the dataset? • Is the public satisfied with the contents of the dataset?	• The public's data requirements are determined and are available • The identified data requirements correspond with the purpose of the dataset • The public is satisfied with the publication of the dataset	• The public's data requirements have not been determined or are not made available • The identified data requirements do not correspond with the purpose of the dataset • The public is dissatisfied with the publication of the dataset or were not consulted

Table 39: Indicators specifying whether the dataset meets the needs of the data users

Element 2: Does the dataset contain beneficial information for the data user?

Once it has been identified if the dataset in its entirety is relevant to the data user, the contents of the dataset must be examined to identify if there are unnecessary and irrelevant details included within the published dataset that have no value to the data user.

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Does the dataset contain unnecessary/superfluous details not required by the data user?	• All sub-components of the dataset must be examined to determine the granular subunits are relevant to the data user	• The dataset presents only relevant and applicable details to the data user	• The dataset presents multiple irrelevant details to the data user

Table 40: Indicators specifying whether the dataset contains information beneficial to the data user

7.2.5. Dimension 6: Conciseness

The conciseness of the data presented to the public is a data quality factor based on how the data product is formed. Although conciseness also refers to the limitation of superfluous and unrelated information, it differs from applicability as the nature of this limitation on irrelevant information also relates to how precisely the information is provided to the user (see section 4.3).

Element 1: Is the data concisely presented to the public?

To determine what data components are relevant for the data user, a clear understanding must be made of which types of information are beneficial to the public. Where the dataset provides information beyond the scope of the public's interests, these components will be found to be irrelevant. This is applicable in terms of the granularity of the data as well as the range of fields included within the public dataset.

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Is the data to the point?	• Does the data contain unnecessary elements for the data user? • Does the data provide a sufficient level of granularity for the data user?	• All data provided is relevant to the data user • Data is provided to the most appropriate and consumable level of detail	• Some data provided is irrelevant to the data user • Data is provided at a too granular level of detail

Table 41: Indicators specifying whether the dataset is concisely presented to the data user

7.2.6. Dimension 7: Consistency

The discussion pertaining to the consistency of data relates to the semantic and structural forms of consistency. The published data needs to present concepts and values which are uniformly expressed and verifiable. Semantic consistency refers to the rules and definitions which are applied in the production of the dataset whilst, structural consistency describes the actual representation of data values (see section 4.3).

Element 1: Is the data semantically consistent?

Semantic consistency refers to the uniform application of rules and definitions when producing a particular dataset. Such rules should follow common practices adopted within the organisation or follow from international best practices. Where a change in approach is adopted, the metadata must clearly state how such rules were altered and what the new basis for the application is. There is a need for common methodologies and data processing steps which can be used to assess how consistently rules and definitions were applied (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are data rules and definitions applied consistently to the dataset?	• Does the data provider follow a common methodology (international standard) when producing the dataset? • Does the metadata describe changes to the methodology and/or rules, definitions that are used?	• The data provider utilises a common international standard with respect to all contained definitions and rules • The metadata provides reasons for all changes made to rules and definitions	• The data provider differs from international standards with respect to some or all contained definitions and rules • The metadata provides no or limited reasoning for the changes made to rules and definitions

Table 42: Indicators specifying whether the dataset applies definitions and rules consistently

Element 2: Is the data structurally consistent?

In addition to examining the data rules and definitions, the actual data values need to be assessed for consistency. Where data is processed within an organisation, the data provider must guard against irregular manipulations which distort the reported value. From the data user’s perspective, however, such internal manipulations within the organisation are difficult to be judged by someone outside of the production system. Therefore, where similar data values are released to the public by an organisation, the reporting of such values should be consistent. Where discrepancies emerge, the accompanying metadata should describe where such differences occur and the reasons for the change in value. This is especially significant in the re-release of only datasets that are updated with figures pertaining to the new time-period. In these circumstances, the data provider should include a proviso which explains the changes in reported values (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are data values consistently reported to the public?	• Are data values over time are reported consistently? • Does the data provider reasons for changes in data values?	• Data values are reported consistently • The metadata describes the rationale for update prior values reported	• Data values across data releases differ • The metadata does not provide reasons for the changing data value

Table 43: Indicators specifying whether the dataset reports data values consistently

7.2.7. Dimension 8: Currency/Timeliness

Currency and timeliness pertains to the amount of time that passes from the point the data is collected to the point that it is reported. The sooner the data is made available the more current the information is. Often with public socio-economic data, the currency of the information is a crucial factor determining the relevance of the data as conditions within communities change frequently (see section 4.3). Furthermore, where a data collection has been conducted and is expected by the public, the punctuality of the data release is a factor which affects data users. Generally, there are standards pertaining to the time period between data capture and data release (see section 4.5.1).

Element 1: Is the dataset punctually released?

In assessing the punctuality of the data release, the time between the data collection and data release must be determined. This time difference is assessed in terms of recommended timeframes and international best practices.

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Is the dataset released in a punctual manner?	• Is the average time between the end of the data collection and the data release within recommended time-frames? • Does the organisation release a report on the time	• The time take to release the dataset is within international standards or reasons are provided for the late release of data • An accompanying time frame is released to the public to ensure they	• The time take to release the dataset is beyond international standards • No time frame report is released or the report is released and/or the organisation does not follow the

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
	frame for data releases?	are aware of when to expect the release	prescribed time-frame

Table 44: Indicators specifying whether the dataset has been punctually released

7.2.8. Dimension 9: Accessibility

Following from the infrastructure related analysis of data quality factors, the accessibility of the provided data is a key factor which impacts how useful the data production is. Data should be reachable, obtainable and retrievable. Access rights to data can limit the impact the data has. Often in public institutions, data access is controlled such that institutions can have a record of what, why and how frequently their data is used. However, such controls often require a human interface which can introduce delays to data provision. In addition, certain granularities of data need to be protected to ensure the anonymity of the survey subjects (see section 4.3). However, if only highly aggregated information is released, the detailed information which is valuable to data users becomes inaccessible. The delivery of data to data users should be controlled, the data and metadata must be equally accessible and the data should be presented in a meaningful and useful manner (see section 4.5.4).

Element 1: Are data access controls too stringent?

Where datasets are protected, clear guidelines need to be offered to the public regarding how to access the published information. Requests for registration or access provision and the terms for such access need to be clearly indicated by the provider. Once such requests are made, the data provider must be responsive to provide the data-user feedback to their request. When granting access to individuals, the data provider must ensure that the access controls are correctly applied and ensure those that require access are provided access (see section 4.3).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are the applicable data users granted an appropriate level of data access?	• Are the access rights clearly and uniformly applied ensuring that the correct individuals are granted access? • Can users easily find directions on how to access the required information? • Does the data provider respond to data requests in line with international	• Data access is uniformly granted following the data security policies • Guidance for requesting data is clearly provided and the steps to do so are simple to follow • The data provider responds to request in a timely manner in line with	• Data access is not uniformly applied • Guidance for requesting data is not provided or is difficult to find • The data provider does not respond to request in a timely manner

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
	standards on responsiveness?	international standards	

Table 45: Indicators specifying whether the data provider grants access to the correct members of the public

Element 2: Is a suitable granularity of data accessible to the public?

Access to individual datasets can be secured as well as access to particular fields or particular categories of data values where necessary. Where such controls are in place, the public must be informed of which fields and categories are accessible (see section 4.3).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are access controls applied to the correct fields and categories?	• Are access controls applied to the correct selection of fields and or categories? • Are adequate levels of anonymity applied ensuring that the user's needs for data are considered? • Does the metadata describe the levels of detail that are accessible?	• Access controls are applied correctly in accordance with the appropriate selection of fields and or categories • Anonymity of data is applied to the most appropriate level considering the user's needs for data • The level of detail of data access available is documented within the metadata	• Access controls are in correctly applied across fields and/or categories • Anonymity of data is wrongly applied • The level of detail of data access available is not documented within the metadata

Table 46: Indicators specifying whether the access controls are applied to the correct level of data granularity

Element 3: Is the data and metadata provided in a meaningful and useful manner?

Data providers need to clearly define how they intend to disseminate their data to the public. The adopted strategy also needs to be shared with the public for them to understand how to access such information. Furthermore, the strategy must describe the data release schedule and the manner in which data can be retrieved. This strategy must also be maintained and updated based on the changing needs of the data users. The organisation must therefore also be cognisant of the data user's needs for accessing data

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Is data and metadata provided in a useful and meaningful manner?	• Is the data dissemination strategy shared with the public? • Is the data release scheduled? • Does the metadata provide guidance on	• The data dissemination strategy is shared with the public and is easily accessible • The data release schedule is available.	• The data dissemination strategy is not shared with the public • The data release schedule is not available.

	how to access and retrieve data using the available mediums?	- The metadata provides steps on how to access and retrieve data	- The metadata does not provide steps on how to access or retrieve data
	- Is the data dissemination strategy frequently updated?	- The data dissemination strategy is updated following changes to user requirements for access	- The data dissemination strategy is not adapted due to changes to user requirements for access
	- Does the data dissemination strategy take heed of the needs of the data users?	- The data dissemination strategy directly refers to needs of data users	

Table 47: Indicators specifying whether the data and metadata are provided meaningfully and usefully to the public

Element 4: Is the channel of data delivery appropriate for the data user?

The final element for data accessibility refers to the use of the particular channel when accessing data by the data user. This refers to the software and hardware requirements and limitations for interacting with the database tools for online databases and the human resources that participate in the data provision chain. Data can be made accessible in a variety of data formats which require particular software applications. For example, statistical datasets may require a statistical tool. It is preferable if the available data that is released is interoperable across various software platforms. The greater the level of flexibility, the greater the uses that can made from the data. Depending on the resources required for accessing data, a certain level of internet bandwidth and memory may be required to access the information. Where data is disbursed using particular installation packages, these packages have particular hardware requirements. Such requirements must be clearly communicated to the data user. In addition, where data is provided via database tools, the appropriate use of such tools to extract the relevant details also needs to be described within the data's metadata which is preferably a detailed user manual that explains how one interacts the system and guides users on the correct process for accessing the data. Lastly, where data provision is not an automated (downloadable) feature but is required to be conducted via direct request by email or phone call, a resource from the organisation should be contactable (see sections 4.3, 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Is the channel of data delivery suitable for the needs of the user?	- Are hardware and software requirements clearly communicated to the data user?	- Hardware and software requirements are communicated to the data user	- Hardware and software requirements are not communicated to the data user

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
	• Have the user's preferred means of accessing data been taken into consideration when developing the data access channel? • Are the disseminated datasets accessible in multiple data formats? • Where database tools are provided, is the mode of access understandable and simple to operate? • Are user manuals provided to aid data access to database tools? • Where external staff resources need to be contacted to request dataset, are such staff reachable?	• Data users are contacted and their requirements for accessing data has been noted • Data disseminated in multiple data formats. • Where database tools are provided, the mode of access is simple to operate. • User manuals provided to aid data access to database tools • Where external staff resources need to be contacted for data requests they are available and responsive	• Data users are not considered • Data is not disseminated in multiple data formats. • The database tools provided are difficult to operate. • User manuals are not provided to access to the database tools • Where external staff resources need to be contacted for data requests they are not contactable or responsive

Table 48: Indicators specifying whether the data delivery channel meets the needs of the public

7.3. Computational Dimensions of Data Quality

The Computational aspects of data quality generally refer to the physical system related decisions that need to be made. The majority of these aspects related to data quality do not impact the data user. Two issues, viz., Integrity and Traceability, emerge and have a direct bearing on the data user as well as their means to assess their particular state. This is described in the following section.

7.3.1. Dimension 10: Integrity

A key factor when assessing data integrity is to assess the strength of the data quality constraints applied within the data production process. By understanding how the data quality has been maintained assists the data user in trusting the published data (see section 4.4). Furthermore, the statistical and methodological transformations applied to the data needs to follow international best practices. These practices need to be well documented within the data's published metadata (see section 4.5.4).

Element 1: Are data quality constraints suitable for the needs of the data user?

When the data user assesses the data quality constraints, they can only review the processes described within the supporting metadata. These constraints must capture the range of data rules applied to the dataset and cover the full range of statistical techniques used to adequately present the dataset to the public, in a manner that is representative of its targeted population group. Furthermore, where breaches occur, the organisation must follow a process to address the breach in quality. The management processes that are undertaken need to be described within the supporting metadata as well (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are the data quality constraints applied within the data system sufficient to manage data integrity?	• Do the data quality constraints follow international best practices? • Are the data rules suitably captured within the data quality constraints? • Is the process to address data quality breaches discussed within the metadata?	• The data quality constraints are consistent with international best practices. • The data rules are well captured within the data quality constraints and documented within the metadata. • The process to address data quality breaches are detailed within the metadata.	• The data quality constraints are not consistent with international best practices. • The data rules are not captured within the data quality constraints or are not documented within the metadata. • The process to address data quality breaches are not provided within the metadata

Table 49: Indicators specifying whether the data quality constraints are suitable

Element 2: Does the data provider introduce the most appropriate statistical and methodological approaches?

The data transformation methodologies applied to the data need to follow international best practices. This applies to all transformations that are conducted during the statistical processing of the dataset to ensure that the sample's responses are adequately weighted for example, and that such techniques apply the international best practices. Furthermore, the methodologies that are adopted need to be detailed within the supporting metadata, to allow the data user to understand what transformations took place (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Are the data transformation techniques suitable?	• Are the applied data transformation techniques applied within the bounds of	• The applied data transformation techniques follow international best practices.	• The applied data transformation techniques do not follow international best practices.

	international best practices? • Are the data transformation methodologies detailed within the supporting metadata?	• The data transformation methodologies are detailed within the supporting metadata	• The data transformation methodologies are not detailed within the supporting metadata

Table 50: Indicators specifying whether the data transformation techniques are suitable

7.3.2. Dimension 11: Traceability

In order to prove the lineage of the data, the metadata must include documentation regarding the source of the data and the sequence of changes applied to it, resulting in the final form presented to the user. By understanding how the source information has been adapted the data user gains a sense of trust in knowing that the transformations can be reproduced if needed (see section 4.5.4).

Element 1: Does the data provider track the data source?

It is essential to keep records of the sequence of data transformations and flow of information within an organisation and from outside of the organisation as well. Therefore, the metadata must highlight the data dependencies that enact changes to the published data set. Furthermore, where external data is part of the data sources that feed the data publication, that data source that was used also needs to be vetted through the PDQAF (see section 4.5.4).

Indicator	Focal Issue	High Quality Trait	Poor Quality Trait
Does the organisation track data transformations?	• Does the organisation document the data sources and transformations which are used within the metadata? • Has supporting information been vetted using the public data quality assessment framework?	• The organisation documents the data sources and transformations which are used within the metadata • Supporting data sources have been vetted using the public data quality assessment framework	• The organisation has not documented the data sources and transformations which are used, within the metadata • Supporting data sources have not been vetted using the public data quality assessment framework

Table 51: Indicators specifying whether the dataset is traceable to all data sources

7.4. Summation of the Public Data Quality Assessment Framework

As described within this chapter, using the three perspectives of data quality, viz., the organisational, architectural and the computational have helped to unpack the underlying details required for assessment of each chosen data quality dimensions. Data quality in itself is a very broad field which affects various stakeholders within the production process of a dataset. These challenges need to be managed by organisations to ensure that the right staff members are responsible for performing the necessary data quality tasks to ensure that the net output of the dataset is useful and meaningful to the data user. Therefore, only the issues that directly impact the data user are selected in the production of the PDQAF.

From the organisational perspective, whilst the organisation leaders ensure that the data system is well managed, all the principles, processes and rules they put into place need to be documented and disseminated to the data users. If the metadata is detailed and provides information regarding the standards and international best practices that the organisation adopts, the data user can thereby apply such techniques and produce a similar assessment of the dataset's data quality, as found in this study. Where the metadata is lacking, despite the best implementation of data quality protection processes, it cannot be determined if the data is of suitable standard and therefore it is difficult for data users to find the data trustworthy or useful.

From the architectural perspective various data quality dimensions have been identified. Each dimension bears direct significance on the data user and provides the means to determine if the dataset is comprehensive, accurate, clear, applicable, concise, consistent, current and accessible. In assessing each of these dimensions, specific elements for each dimension are identified and thereafter individual indicators which describe how the dimension should be measured are determined. Similarly, at the computational level, factors of integrity and traceability were identified with accompanying elements and indicators.

As the framework that has been produced is intended to be practically used, specific focal issues and positive/negative traits have been identified which allows one to apply the framework directly to each identified dataset and supporting metadata publication. Where the IMF DQAF provides the granular breakdown of dimensions into elements, indicators, focal issues and key points, this study adapts the key points into positive and negative traits using which allows one to easily determine if a dataset has greater or fewer numbers of high quality traits compared to poor quality traits. These traits can be used in the form of a checklist to identify how well a dataset performs in terms of the identified dimensions of data quality.

In the following chapter, individual datasets are identified following from the identified functionings determined with in the school dropout Capability Set as outlined within Chapter

6. Each of the selected datasets per functioning are thereafter analysed in terms of the dimensions discussed within the PDQAF.

Part III

Findings

Chapter 8

Application and findings from the School Dropout Capability Assessment Framework and the Public Data Quality Assessment Framework

The application of the constructed School Dropout Capability Assessment Framework (SDCAF) and the Public Data Quality Assessment Framework (PDQAF) requires the selection of publicly released data by the national statistical agents, education ministries and other research bodies or data collectors of Brazil, India and South Africa, which are relevant to the functionings, themes and sub-themes identified within the SDCAF. Specific questions or variables (depending on the type of dataset selected) in the released datasets are selected where the premise of the question follows the core concerns of the particular thematic area as discussed in Chapter 6. Once datasets and questions within these datasets are identified, the entire released data collection instrument is assessed by applying the PDQAF. This application of the PDQAF is used to answer research Question 4 of this study which requires the identification of data concerns which may impede the identification of relevant datasets which inform the assessment of the SDCAF.

Using the 11 dimensions of data quality, one is able to determine which particular type of data concerns impedes its use by data users. The data quality elements, described in terms of indicators and positive or negative focal issues are used to determine whether the dataset in its entirety (not just the particular questions) is of a suitable quality for data use. An aggregation of the positive and negative focal points informs the assessment of data quality, whereby each dataset is scored in terms of how strongly the dataset performs per data quality dimension based on these positive and negative focal issues. The aggregate score per dataset informs the data user of the data quality and is crucial in testing whether relevant published datasets can be used by the public. By linking individual data questions or variables of datasets to the SDCAF thematic areas, the scores of the individual dataset inform whether the available data is acceptable for reporting on such themes. It is recommended that datasets of poor data quality are avoided.

Figure 6 below, presents a summary of the assessment of each data collection instrument that was found relevant for this study, using the PDQAF. Due to the greater number of datasets that were identified in South Africa, the majority of the poorly assessed surveys were South African. However, data produced by Statistics South Africa and the National Income Dynamics Study were also found to be the strongest performing datasets across the three countries, in terms of data quality. Interestingly these are produced by data providers that closely follow the prescripts of South Africa's SASQAF.

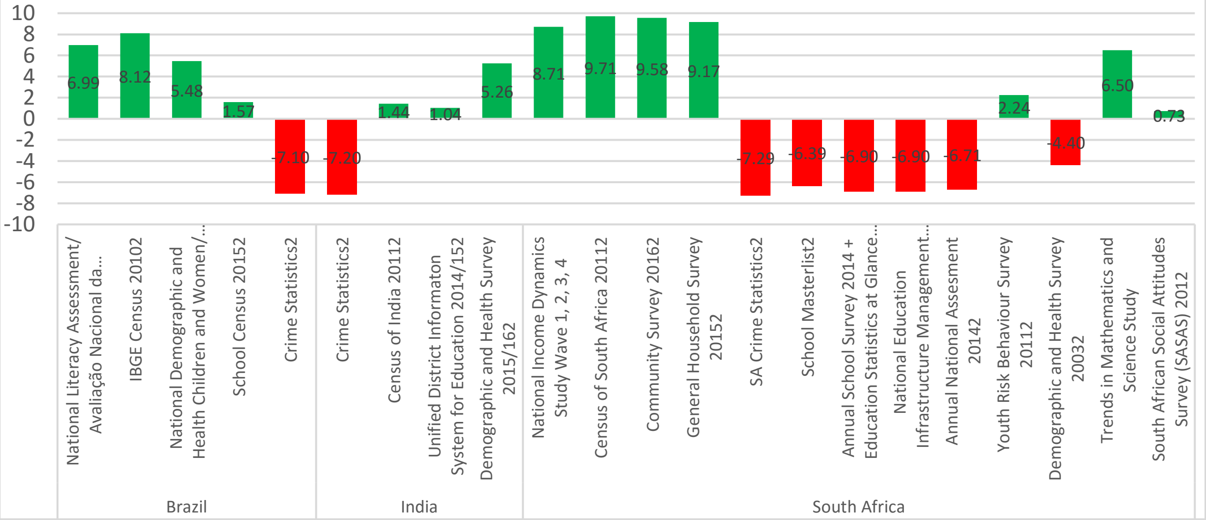


Figure 6: Data quality assessment summary of data collection instruments conducted in Brazil, India and South Africa

In order to test the application of the framework, the first functioning included in the SDCAF (Learner's Physical Well-Being) is selected to prove whether the concepts included within the SDCAF and PDQAF can be used to test the identified functionings identified in the SDCAF Capability Set. Whilst using this approach, this study can pass judgement on the suitability of data to assess the School Dropout Phenomenon in its entirety and also advise how such an assessment can be made, if the tests applied are extended to the remaining functionings of the SDCAF. Furthermore, the approach that was followed can be adapted to be applied to other social phenomena affecting the BRICS.

As each identified data collection instrument has been tested using the PDQAF the detailed review of each dataset is available in Appendix 1 and Appendix 2. Using the combination of the SDCAF and the PDQAF, each theme related to the learner's valuing of physical well-being is assessed in this Chapter. Where no datasets are identified for a sub-theme, these gaps are highlighted per country. It is important to note that even data gaps are valuable findings, as such gaps signal a data collection need that has not been met by the identified data providers. Furthermore, the assessment of the PDQAF identifies the level of data quality that is associated with selected question or identified variable.

Once the data gaps and level of data quality per thematic area are identified, a few examples of data reports are produced to highlight some of the functionality of the available data. Following this presentation of data examples, the selected data variables are assessed to determine whether the conceptual concerns per thematic area have been adequately represented.

8.1. Public Data Quality Assessment Framework: Being Physically Healthy

In answering research question one of this study, datasets relevant to the school dropout phenomenon were identified through a non-exhaustive review of data produced by data providers from Brazil, India and South Africa. Within these datasets, pertinent data questions or variables that describe the functionings of the SDCAF were selected. Each data collection instrument that was selected was analysed using the PDQAF (discussed in detail in Chapter 7). The findings of this assessment present an interesting summary of data quality issues affecting the three countries. From this assessment, one can identify that each country has produced data collection instruments with varied degree of data quality.

In general data produced by statistical agents or international bodies tend to score higher than those of government departments who may not be as well skilled or trained in producing quality data. In these organisations, such as the Police Services for example, producing quality data is not their primary concern and therefore the organisation may not have built the institutional capacity to ensure data quality concerns are suitably managed or fully incorporated in the production process. In these organisations it is found that the data that is released is not accompanied by any supporting metadata, resulting in poor data quality scores based on the PDQAF assessment. From the review of the identified datasets in this study, it is clear that organised and detailed metadata is the primary means that a data provider has to communicate to the data user to what degree of data quality that they have access to when using the available dataset. In the absence of published metadata, the public is forced to assume that no data quality controls are in place, despite scenarios where such practices may be followed by the data provider. The assessment that follows is based on the published metadata and datasets that were identified as relevant to the learner being physically well. For each theme and sub-theme of the physical wellbeing of learners, the data collection instruments are reviewed to identify commonly themed questions and thereafter are assessed using the PDQAF. For purposes of this assessment, only the latest available dataset per theme has been assessed.

Table 52, below, provides an extract of the PDQAF assessment applied to surveys from Brazil, India and South Africa. This extract provides an example of how the PDQAF was applied, specifically focussing on the Metadata dimension's and in particular the data quality element "Does the metadata convey the contents of the dataset?" The indicators and focal issues, linked to this particular data quality element provide the means to assess whether the surveys have addressed each relevant facet. For a detailed review of the PDQAF for each identified data collection instrument, consult Appendix 2 and Appendix 3.

Element	**Indicator**	**Focal Issue**	**Brazil: IBGE Census 2010**	**India: DISE 2014-15**	**South Africa: Annual National Assessment 2014**
Does the metadata convey the contents of the dataset?	Are concepts and definitions and classification provided within the metadata to describe the underlying data?	Are definitions made available within the metadata?	Concepts and definitions are discussed in detail in the Census results documentation	Concepts and terms are defined within an Educational Planning Guidebook	Definitions are not discussed within the ANA report
		Are deviations from standards reported?	All concepts are well defined throughout the various pieces of documentation	Effort is being made to align the data collection to international best practices, but the deviations from best practices are not referred. Only copies of signed agreements are published on the DISE website.	No standards are identified
	Is the metadata up to date?	Is a document register regularly maintained in line with the data collection strategy?	The documentation is provided for each census	The guidebook is outdated. Was produced in 2003	A data register is not provided

Element	Indicator	Focal Issue	Brazil: IBGE Census 2010	India: DISE 2014-15	South Africa: Annual National Assessment 2014
	Are all tables and fields defined?	Is the purpose and definition of each table and field within the dataset reported on?	Data is provided in various table structures on the IBGE Website at different levels of aggregation, not every restructure of census data is accompanied by documentation with definitions of the terms that are used.	The metadata discusses all the concepts in general. Individual surveys, tables and fields are not described	Tables and fields are not defined
			Fields used in the census are consistent with their use in Household and other demographic surveys of Brazil	Fields in tables follow from descriptions in the guide book	Tables and fields are not defined
	Is the geographic distribution clearly defined?	Is the geographic level of data granularity described within the metadata?	The granularity is documented; information is provided to sub-district level via the IBGE website	The Geographic granularity is not described within the handbook	No metadata was provided
			Geographic boundaries are well documented	The Geographic granularity is not described within the handbook	Definitions are not discussed within the ANA report

Table 52: Extract from the PDQAF assessment in Brazil, India and South Africa

In assessing the data that describes learner's physical well-being, 2 datasets were found in Brazil, 3 datasets in India and 8 datasets from South Africa. In applying the PDQAF, the majority of the datasets achieved a positive data quality result apart from datasets from the South African Ministry of Basic Education and South Africa's Demographic and Health Survey of 2003. In addition, although most datasets were positively assessed, not every dataset achieved a high data quality score. Therefore, data concerns were exposed. As discussed in section 5.5**Error! Reference source not found.**, the aggregation of positive and negative focal i ssues is summarised to produce the data quality scores as described in Figure 7. As the data quality score is rebased to provide scores between 10 and -10, datasets which attained a score above the value of 5, tended to display mostly positive data quality traits. These datasets are rated as ‘Quality data’ using an adaption of Statistics South Africa’s rating as described in section 4.5.4. Similarly, datasets which scored between 0 and 5 are accessed as ‘Acceptable data’, those that scored between -5 and 0 were assessed as ‘Questionable data’ and lastly the datasets that scored between -10 and -5 were rated as ‘Poor data.’

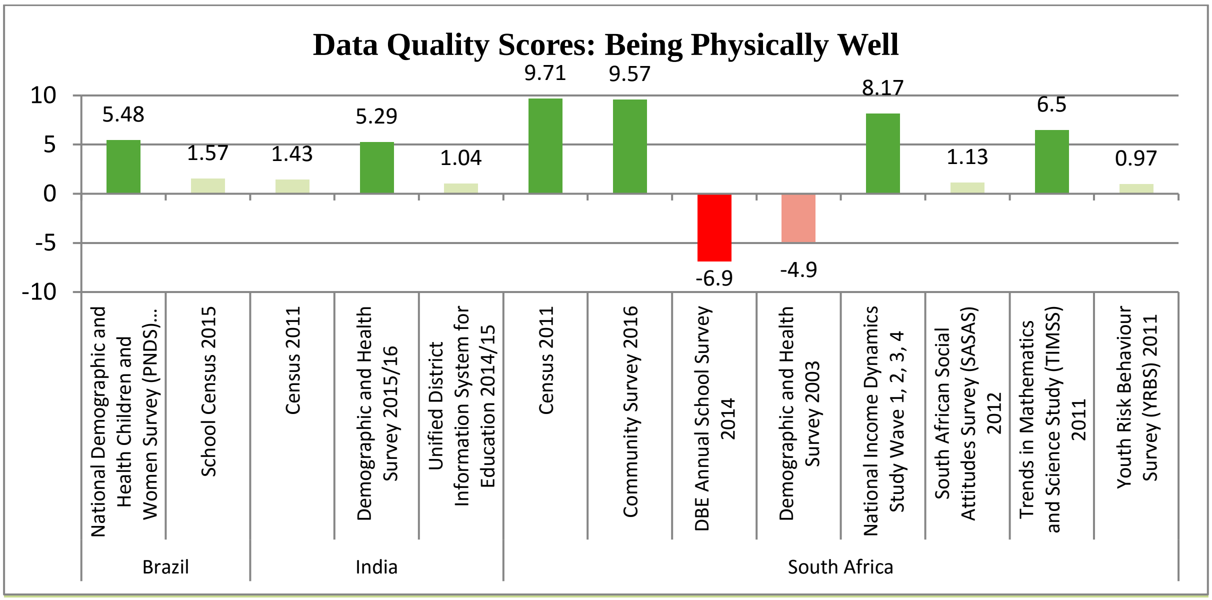


Figure 7: Data Quality Scores related to the learner being physically well

8.1.1. Brazilian Data Sources

In Brazil, 2 data surveys were identified, viz., the National Demographic and Health Children and Women Survey (English translation of the title) of 2006 and the School Census of 2015 as described in the table below.

Data Provider	Collection Instrument	Score
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)	School Census 2015	1.57
Ministry of Health	National Demographic and Health Children and Women/Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher (PNDS-2006)	5.48

Table 53: Data Collection Instruments in Brazil applicable to being physically well

National Demographic and Health Children and Women Survey 2006 (Score: 5.48 – Quality data)

The National Demographic and Health, Children and Women Survey was conducted in 2006 (also referred to by the Portuguese acronym PNDS 2006) and is managed by the Brazilian Ministry of Health. The dataset attained a score of 5.48 and is described as Quality data, using the rating scale discussed above. The survey performed strongly in terms of data clarity, comprehensiveness, conciseness and consistency.

As the survey is also supported by the International Demographic and Health Survey Programme where the questionnaire structure and data quality practices followed by the Health Ministry are based on internationally recognised standards which are recognised as important positive focal points within the PDQAF. The survey was found to be weakest in the areas of data accessibility, applicability and currency. As found in the analysis of other Brazilian data collections, the detail provided in the published metadata is critical to attain a higher data quality assessment score. For this particular survey, the useful metadata dimension achieved a score of 0.5 due to the metadata not reporting on all the structural details of the published data, not providing the geographic detail provided within the survey, lacking documentation on the measuring of data quality and due to the metadata's exclusion of hardware or software requirements for accessing the data. These factors limit the metadata's strength in terms of the contents of the dataset, the context of the dataset and most severely in terms of the metadata's explanation of the structure of the dataset. However, the metadata was found to be strong in terms of other focal issues such as the definitions used within the data, the linkage of the definitions to international standards, the manner in which data quality is processed by the organisation, the data quality controls that were instituted, the comprehensiveness of the data collection methodology and the inclusion of a detailed supporting findings reports.

Whilst the PNDS provides a wide range of crucial data variables and follows internationally recognised standards, the shortcomings of the survey relate to the limitations found in the metadata regarding the geographic structure and the lack of a detailed codebook that provides the data user information regarding how to access and interact with the released datasets. These oversights together with the exclusion within the documentation of how data breaches were addressed tended to reduce the data quality score that was achieved.

Brazil School Census 2015 (Score: 1.57 – Acceptable data)

The School Census of Brazil in 2015 as coordinated by Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (National Institute of Educational Studies and Research in English) provides many questions relevant to the various functionings valued by learners in terms of the choices they make in regards to dropping out of school. The survey is conducted annually and pools the efforts of the public and private schools as well as state and municipal departments of education. The survey is very broad and targets all learners and school professionals from all schools in the country.

The Brazil School Census suffers in many of the data quality dimensions, which results in the survey attaining a rating of ‘Acceptable Data’. The positively assessed dimensions are the clarity of the dataset, how comprehensive the data is and how concisely the content of the information is communicated to the data user. In terms of clarity, it was found that the survey was consistent in the manner it used terms and definitions in the broad survey. Ensuring that the data, metadata and published results are kept consistent supports a common interpretation and understanding of the data. In terms of comprehensiveness, the survey was found to adequately represent all points raised within the metadata and in the actual data that was published. This included all statistical components discussed, the physical structure of the data, the manner in which the dataset presents items discussed as part of the scope, definitions and classifications of content as well as the manner in which the dataset represents the rules discussed within the metadata. In terms of conciseness, the dataset was found to be accurate with respect to the range of fields provided and with respect to the depth of granularity that is provided.

In assessing all the data quality dimensions, the degree of useful detail provided within the metadata drives the data quality scores achieved for the various data quality dimensions. Whilst INEP makes an effort to discuss the structure and definitions of the very broad questionnaire that is rolled out annually, the organisation does not document the data quality practices that may be followed. In the absence of such details, the user must be cautious and assume that such data quality practices were not performed. If greater detail related to the data quality factors are documented within the metadata, a higher score may be achieved.

8.1.2. Indian Data Sources

In India, 3 data collections were identified pertaining to the physical well-being of the learner, viz., the data sourced by the Unified District Information System for Education (U-DISE), the Census of India and the Demographic and Health Survey as detailed in the table below.

Data Provider	Collection Instrument	Score
Office of the Registrar General & Census Commissioner, India (ORGI)	Census of India 2011	0.68

National University of Educational Planning and Administration (NUEPA)	Unified District Information System for Education (U-DISE) 2014/15	1.04
DHS Program	Demographic and Health Survey 2005/6	5.26

Table 54: Assessment of Indian Data Collection Instruments related to being physically well

Unified District Information System for Education 2014/15 (Score: 1.04 – Acceptable data)

The Unified District Information System for Education (U-DISE) is managed by the National University of Educational Planning and Administration and provides a statistical publication each year on the performance of the educational ministries in the country. As found with other data collection instruments, the low data quality score of 1.04 that was achieved is a consequence of the level of detail provided within the released metadata.

With respect to the content covered in the metadata, the dataset's terms and concepts are detailed within the released 'Education Planning Handbook'. However, the handbook was last updated in 2003 and is not current. In addition, the individual surveys or data collections conducted by the Ministry are not referred to in the handbook, and neither are details of the specific fields and tables documented. With regards to the context of the metadata that is provided, the data quality practices are discussed to a limited extent, where they referred to data sampling techniques that were followed. However, the manner in which U-DISE measures data quality is not discussed extensively.

In terms of how data quality concerns are processed, it is noted that U-DISE releases reports on the random sampling that is conducted to discourage individual school districts from incorrectly capturing data values. Schools have an incentive to do so as their funding of is linked to the reported number of learners, educators and facilities. In addition, the metadata holds some insights about U-DISE's data collection methodology, whereby the manner in which the ministry targets every school is documented. Thereafter 5% of schools are surveyed to assess their data correctness. However, these discussions within the metadata are quite generalised in nature and do not refer to the specific processes that are followed when individual questionnaires are circulated to schools to test the data correctness.

One of the primary concerns with the U-DISE data collection, apart from the metadata weaknesses is the limited release of data in usable formats. By providing multiple individual datasets in PDF format, the format is inflexible for data analysis and the multitude of reports makes it difficult for the organisation to provide supporting documentation regarding the use and purpose of each spreadsheet. Larger datasets with a single set of supporting documentation that describes the data collection effort as a whole would be more useful to the data user, in addition to being simpler to manage for NUEPA.

It is recommended that greater emphasis is given to managing data quality constraints during data collection and thereafter documenting such practices in the metadata.

Census of India 2011 (Score: 0.68 – Acceptable data)

The Census of India is conducted every ten years and is managed by the Office of the Registrar General & Census Commissioner, India (ORGI). The Census provides data on the financial security of the learners, the home learning environment and the physical health of the learners. The detail and depth of the information makes it very attractive as a data resource but the collection infrequency limits the currency of the information from the perspective of the data user.

In review of each data quality dimension, the Census scores both positively and negatively for each dimension in terms of the individual focal issues. The negatives are largely offset by the positives resulting in scores of 0 for each dimension apart from data comprehensiveness and useful metadata which are slightly positive when all focal issues are summarised across those dimensions. On closer inspection of the metadata usefulness dimension, the bulk of the negative focal issues are related to the context of the metadata. The content of the metadata, the concepts, fields, tables and geographical boundaries are discussed in the supporting documentation. However, when looking at the contextual issues within the metadata, it is found that the metadata does not describe in sufficient detail the data quality practices followed by the Office of the Registrar General and Census Commissioner. Although the metadata alludes to a pilot study that was set up to investigate the pressing concerns for the census, the metadata does not provide further specifics about data quality management, statistical assessments or a discussion on how data quality concerns were addressed during collection.

Furthermore, the metadata does not provide sufficient detail to allow the data user to assess whether the data collection process had comprehensively reached the targeted population. While it is understood that reaching the entire population of India in a single survey is difficult, the published metadata does not provide any information regarding the response rate of collection. Furthermore, the questionnaire structure is not discussed and the metadata does not provide any references to international norms and standards which may have been followed to guide how the census operations were carried out. On the positive side, a detailed findings report does accompany the release of the data and the documentation details the necessary concepts which are consistently applied. In terms of the structural issues affecting the metadata it is found that the physical structures of the census are documented.

From a data accessibility perspective, the data dissemination strategy that was followed entailed providing only static excel spreadsheets which limits analysis such as multivariate cross-tabulations not offered within the static excel spreadsheets. In terms of the detail provided, for each file the fields provided are based on the fields that are mentioned within the metadata code book. The primary limitation of the available data is the inaccessibility of analysing the database in its entirety. To access greater detail, users needed to find Census University Workstations which are located at a selection of universities across the country. At these workstations detailed databases are installed allowing users the opportunity to more closely examine the census results. However, these installations are quite restrictive as one’s results could only be extracted by physical printout with no digital downloads of the information allowed. If a user requires a particular cross-tabulation of data not provided in the static spreadsheets made available online, a request must be submitted to the ORGI staff.

The major data quality impediments affecting the Census of India (2011) relate to the lack of detail within the metadata discussing the data quality processes such as response rates or imputation rates as well as the restrictive data dissemination policy. In addition, the lack of online data tools or the release of granular data downloads limits the public from performing detailed analysis. The only users that can perform such tasks are those close to the selected universities which house census workstations. Even in these situations, data is not freely accessible as no data downloads from the workstation are permitted.

Demographic and Health Survey 2005/6 (Score: 5.26 – Quality data)

The Demographic and Health Survey (DHS) of 2005/6 for India is managed by the international Demographic and Health Survey Program which is funded by the United States Agency for International Development (USAID). The survey provides nationally representative data pertaining to trends on health and population matters.

The DHS of 2005/06 attained a data quality score of 5.26 which supports a rating of ‘Quality data.’ The dataset performs strongly in terms of data clarity, comprehensiveness, conciseness and consistency. However, the dataset’s quality score could be improved if certain additional items were included in the supporting documentation. This is reflected in the useful metadata score that suffered due to gaps in the metadata such as missing documentation of the provided data tables and the lack of detail about the geographic granularity of the data. Although data on India is meant to be nationally representative, the metadata at no point confirms how granular the data is representative to. A specific mention of such details would reduce the assumptions that data users are forced to make. Although, sampling plans are provided and data collection scenarios are provided to help train field workers when conducting the survey, the organisation fails to explicitly define data quality and state how it is measured and thereafter managed.

The weaknesses of the survey primarily relate to the lack of detail pertaining to the released tables and the lack of a clear data dissemination strategy. Due to these concerns, various data elements were scored negatively despite the survey holding many positive traits based on the manner that it follows international best practices. The survey is supposedly nationally representative but due to a lack of published response rates this cannot be confirmed.

8.1.3. South African Data Sources

Of the various datasets found relevant in terms of school dropout factors, 8 of these datasets were found to provide some data in terms of the learner's physical well-being. These include the Census, the Community Survey, the Annual School Survey, the Demographic and Health survey, the National Income Dynamics Study, the South African Social Attitudes Survey, the Trends in Mathematics and Science Study and the Youth Risk Behaviour Study. The scores attained by each of these surveys are discussed in the table below.

Data Provider	Collection Instrument	Score
Southern Africa Labour and Development Research Unit, University of Cape Town	National Income Dynamics Study Wave 1, 2, 3, 4	8.71
Statistics South Africa	Census of South Africa 2011	9.71
	Community Survey 2016	9.58
South African Police Services	SA Crime Statistics	-7.29
South African Department of Basic Education	Annual School Survey 2014 + Education Statistics at Glance Publication 2014	-6.90
SA Medical Research Council	Youth Risk Behaviour Survey 2011	0.97
	Demographic and Health Survey 20032	-4.90
Human Sciences Research Council	Trends in Mathematics and Science Study	6.50
	South African Social Attitudes Survey (SASAS) 2012	1.13

Table 55: Assessment of South African Data Collection Instruments applicable to being physically well

Census of South Africa 2011 (Score: 9.71 – Quality data) & Community Survey 2016 (Score: 9.58 – Quality data)

Statistics South Africa produces 2 surveys which are relevant in the assessment of learner dropouts. This includes Census 2011 and the recently released Community Survey 2016. The surveys produced by Statistics South Africa predominantly report on information pertaining to population demographics. Using the person age variable, the demographics released can be filtered to the school going age to determine the trends affecting the learner population.

The Statistics South Africa data collections have performed the strongest in terms of the data quality assessment. The Census of 2011 and the Community Survey of 2016 scored values of 9.7 and 9.58 out of a possible 10 points respectively, thus achieving a rating of 'Quality data'. To achieve this particular score, both surveys scored strongly across each dimension where the assessment was applicable. Where the Census is from 2011, the Community Survey data collection was only recently completed at the time of analysis of this particular survey and therefore the detailed release of data was not available at this particular moment.

The data quality strengths of the various dimensions follow from the strength of the reported metadata which accompanies the surveys. Statistics South Africa performs particularly strongly in this regard as the organisation follows the SASQAF which was instrumental in defining the focal issues used in the PDQAF. Therefore, in terms of the metadata's conveyance of the content, context and structure of the metadata, the documentation for both of the surveys were adequate as it supported each of the required elements and indicators. The concepts and terms used are well detailed and are accompanied by data registers which provide information about the relevant tables and fields made available per survey. In terms of the metadata's description of the context to which the dataset is situated within, a weakness across both dataset is the lack of measurements provided by Statistics South Africa on the level of data quality attained. However, the metadata does clearly articulate the position the organisation takes in applying the SASQAF principles.

The strength of the Statistics South Africa datasets compared to the Brazilian Census or the National Income and Dynamics Study is the emphasis within the metadata about the importance of defining data quality using the SASQAF principles. Statistics South Africa's datasets are set apart from other data providers primarily in terms of the manner in which the organisation centralises the application of the SASQAF in all aspects of data collection. With respect to the data structures, the Census performs well in the manner the documentation describes the makeup of the individual tables and fields. However, as the Community Survey has not been released as yet, the detailed data is not available for download and therefore supporting documentation of the individual table structures is not released and therefore cannot be assessed.

In review of the surveys conducted by Statistics South Africa, it is noted that the degree to which the SASQAF is implemented and discussed in detail, in the published metadata, is an example to the other members of BRICS, as their surveys emphasize data quality during data collection and production. By ensuring that these processes are well documented, Statistics South Africa provides the data user a sense of surety and trust that their data is robust and reliable.

Annual School Survey 2014 (Score: -6.90 – Poor data)

The Department of Basic Education (DBE) is responsible for a variety of country wide collections using surveys and registers targeting all schools in the country including the Annual School Survey. In review of the individual data quality scores, one finds that each factor apart from data conciseness is rated negatively. With the majority of dimensions scoring quite poorly, the root of the data quality concerns begins with the lack of any released supporting metadata by the department. With the lack of metadata, the assessment of the content, context and structure of the metadata is impossible and therefore the assessment of each indicator and sub-element of data quality is found to be negative. Importantly, there are no details communicated to the data user regarding the definitions of terms or any information regarding how the department defines and manages data quality. This lack of available documentation for the data user to peruse is the primary cause of the poor scores for each following dimension.

Data accessibility is found to be quite a challenge with respect to the DBE data collections. Whilst some of the Annual School Survey data is published within the ‘Education at a Glance’ publication, the bulk of the DBE data is strongly protected and is not published online. Data requests to the department need to be made and responses are not always guaranteed. Data from these data collections is often referred to in documentation from the department and the data user is forced to consume the ‘2nd hand analysis’ of the data and make inferences on the data based on such information.

The DBE may actually follow some data quality controls when collecting this data, but the efforts made by the department to do so are not documented and communicated to the public and therefore the user must assume no quality processes are in place. Thus the DBE data collections are assumed to be unreliable due to the lack of information provided to the public.

Demographic and Health Survey 2003 (Score: -4.90 – Questionable data)

The South African Demographic and Health Survey (DHS) of 2003 was released in 2007 by the South African Medical Research Council (MRC) together with the South African National Department of Health. Although the DHS of 2003 follows the international guidelines of the Demographic and Health Survey Programme funded by the United States Agency for International Development (USAID) the metadata makes little mention of the connection to the international body in terms of the various standards and principles that are followed when carrying out the survey. Although the survey results are released by the MRC, the released documentation does not provide any details about the structure of the datasets and the dataset is not actually released. All data is made available within the findings report of the survey. These limitations are reflected in the poor data quality score of the survey (-4.90) as a whole and the low score for metadata usefulness.

The difficulties in assessing the metadata result from the packaging of the metadata within the findings report as an appendix to its release. This results in a lack of detail provided in the report regarding the file layout as well as any discussion on how data quality faults are addressed and also excludes definitions of fields that are used within the findings report. Within the metadata section of the findings report there is a discussion about how field-workers are trained to conduct the survey correctly. In addition, the sampling framework is discussed within the inclusion of the response rates confidence intervals and sampling errors. These details signal that data quality is considered, just not reported effectively.

In review of the survey in its entirety, the lack of available data files for the public to review and work with is a major concern. The discussion within the metadata pertaining to the sampling frame and other statistical calculations that were conducted alludes to data quality practices that were prioritised internally within the MRC. However, these practices are not fully documented. The inclusion of metadata as an appendix to the finding report signals that metadata is not viewed as a priority within the MRC. The data release would benefit from a metadata focused report which pays clear attention to the organisation's requirements for data quality. Without such documentation, it is assumed by the data user that data quality is not a priority for the MRC.

National Income Dynamics Study Wave 1, 2, 3, 4 (Score: 8.71 – Quality data)

The National Income Dynamics Study (NIDS) is a panel study conducted across households in South Africa by the South African Labour and Development Research Unit (SALDRU) at the University of Cape Town. The NIDS utilises a sample that is nationally representative with four waves of the survey that have been conducted since its inception in 2008. As the study follows the panel study methodology, all efforts are made to contact the initial 28000 respondents that were identified in the initial survey. Each wave targets these individuals and tracks their changes over time. In this respect, the survey is costly to manage but produces very valuable and rare data insights across a range of subject areas.

The four waves of the NIDS panel study achieved a data quality score of 8.71 which underpins the great care that has been taken to produce high quality data outputs that the public can trust when reporting on national data trends. The dataset scores highly except in the inapplicable dimension of data traceability. The strong data quality assessment emanates from the detail provided within the metadata which provides a high level of detail describing the content, context and structure of the published dataset. In terms of the content, definitions are clearly presented in an accompanying codebook to the released data for each wave for every released data file, together with well-defined concepts that are included within the documentation.

In terms of the context of the dataset, the data quality practices that are applied are highlighted within the technical documentation which outlines the various processes followed by SALDRU. Furthermore, the documented response rate achieved by the survey is above 90% for each of the waves indicating to the user that the data quality measures were recorded by SALDRU. In addition, the internal practices that are followed to ensure data quality are discussed within the supporting documentation. In terms of metadata comprehensiveness, the statistical checks followed by SALDRU are documented together with information about the sampling frame that was used, the details pertaining to the questionnaire structure and the scope of the particular data collections.

A minor weakness that was highlighted was the documentation alluded to Statistics South Africa practices that were followed but did not identify which practices of the SASQAF were adopted. Lastly, in relation to the structure of the dataset, it is found that the individual data files are well documented and the handbooks provided clearly inform the user how to access the various datasets provided.

The NIDS survey is a promising dataset neither produced by a governmental department nor the national statistics provider. Following guidance from Statistics South Africa and other leading data bodies, SALDRU has produced a high quality dataset producing deep and valuable insights not provided by other data collection processes in the country.

South African Social Attitudes Survey (SASAS) 2012 (Score: 1.13 – Acceptable data)

The South African Social Attitudes Survey (SASAS) conducted in 2012 was managed by the Human Sciences Research Council (HSRC). The survey is conducted annually and is based on a provincially representative sample of 3500 to 7000 individuals aged 16 and older. The survey is used to monitor a set of general demographic, behavioural and attitudinal related questions that are supplemented with additional modules each year based on demand related to topical current affairs issues.

In review of the achieved data quality score of the SASAS 2012, the survey's score reflects a mixture of positive and negative reviews for the data quality dimensions. The overall quality score of 1.13 (ranked as 'Acceptable data') corresponds strongly with the positive and negative assessment of the usefulness of the metadata provided to the public by the HSRC. Whilst the HSRC has made attempts at providing supporting documentation to the public, various short-comings are exposed in terms of the context and content of the metadata provided. These weaknesses affect dimensions such as data integrity, consistency, applicability and accuracy in particular and weaken other data quality dimensions such as accessibility and conciseness. Furthermore, the HSRC data policy of embargoing the release of the most current iterations of the survey by two years severely impedes the currency and therefore relevance of the data release. Furthermore, the HSRC data embargo detracts from the currency of the annual data collection as each annual collection is released at a point when the data may no longer be relevant.

The great value of the HSRC data collection is undermined by the lack of detailed metadata pertaining to the data collection processes followed and documentation of the data quality practices employed. The HSRC however can be commended in terms of their regular collection of data, the accessibility of data products released to the public and the comprehensiveness and clarity of the data that is made available freely each year.

Trends in Mathematics and Science Study (TIMSS) 2011 (Score: 6.50 – Quality data)

The Trends in Mathematics and Science Study (TIMSS) is conducted by the Human Sciences Research Council (HSRC) for South Africa following the international guidelines specified by the International Association for the Evaluation of Educational Achievement (IEA). The survey was conducted in South African in 2003 and 2011. In 2011, the study was conducted in 285 schools reaching 11969 learners across the country. The study targets learners, educators and principals collecting responses on issues faced within the schooling system in addition to the mathematics and science assessment questions posed to grade 9 learners.

The TIMSS 2011 survey scored highly in the data quality assessment achieving a score of 6.5 (recognised as 'Quality data') with a strong performance across dimensions except for applicability and currency. Similar to the analysis of other datasets, the data quality score follows closely from the assessment of the usefulness of the metadata. The reasons for the high performance of the TIMSS study compared to the SASAS survey is that the survey is based on an internationally agreed template with standards that are rigorously tested in numerous environments. Consequently, the definitions of terms and concepts are well detailed and supported with various pieces of documentation. Although the HSRC does not define data quality in the released documentation, various pieces of technical documentation are produced which detail how data quality is protected and ensured. In addition, the confidence interval of the data is described, the quality assurance processes are explained, the sampling framework is comprehensively documented, the questionnaire structure is discussed, the scope of the study is covered and the various concepts, definitions and statistical techniques that are applied are well discussed and referenced within the metadata provided to the user. Ultimately, the content, context and structure of the data files released are highly detailed following the international TIMSS standards leading to the positive scoring across all elements and indicators of the useful metadata dimension.

The TIMSS strong performance signals that non statistical agents can produce data of high quality. Following the data quality requirements identified by the parental TIMSS body, the HSRC was able to produce data that scores highly in a series of dimensions. The data is extremely relevant, though is infrequently produced. Together with the restrictive HSRC data embargoes, the survey loses its impact due to its late release.

Youth Risk Behaviour Survey (YRBS) 2011 (Score: 0.97 – Acceptable data)

The Youth Risk Behaviour Survey (YRBS) was conducted last in 2011 by the South African Medical Research Council (MRC) together with the Human Sciences Research Council (HSRC). The survey updates the results of the previous surveys conducted in 2002 and

2008 in assessing the levels of risky behaviour amongst youth at a provincial level across the country.

The YRBS of 2011 scores highly in some data quality dimensions. The positive, but weak performance in the usefulness of the provided metadata underlines the comparative mediocre data quality score of 0.97. The weakness of the metadata is attributable to a lack of detail describing the structure of the data files produced, as no data file codebook is released to the public, data quality breaches are not discussed, issues of scope are excluded, the metadata lacks references to any guiding standards that are used in setting up the survey and the particular hardware and software requirements for analysing the survey are left out from the documentation all together. These concerns diminish the various positives found in the metadata that relate to details about the sampling framework, the incorporation of a pilot study to test the results and the various validation processes that were applied and documented.

The major concerns in data quality related to the YRBS are the lack of a clear data dissemination policy. Whilst the HSRC assisted the MRC in data capturing, neither organisation has taken responsibility to release the data to the public. In this regard, general data users would fail to assess the survey at all. However, after contacting the principal investigators responsible for the survey, the dataset was accessed and thereafter assessed using the PDQAF. From this assessment, one finds that the lack of a provided codebook and the limitations of a single data file format restricts the usability of the dataset. Care must be taken by the data user to link fields within the SPSS data files to specific questions on the survey. However, in review of the documentation that is provided, it is noted that various quality controls are put in place resulting in a strong performance in terms of accuracy, clarity, consistency and conciseness.

8.1.4. Public Data Quality Assessment - Summary

In review of the related data collection instruments made available in the three countries, it is noted that data collected by national ministries or their supporting agencies or programmes, tend to perform poorly in terms of data quality with some exceptions. The surveys following international guidelines tend to be structured well and provide the data user the necessary details to make their own assessments of the level of data quality per survey. However, where such details were not provided within the metadata, in general, these exclusions led to a poor a data quality rating in other dimensions as well. Notably in Brazil, the National Demographic and Health Children and Women Survey of 2006 performed well. In

India it was the Demographic and Health Survey of 2003 and in South Africa the two Statistics South Africa Surveys, viz., the National Income Dynamics Study and the Trends in Mathematics and Science Study, produced high data quality results. The worst performing dataset across the three countries is the Annual School Survey of South Africa which scores poorly mainly due to the lack of any metadata and the limited data access provided to the public.

8.2. School Dropout Capability Assessment Framework: Functioning – Being Physically Well

Using the data collection instruments itemised in section 8.1 for each country and their associated data quality scores, one can assess the available data quality per sub-theme in relation to the learner's physical well-being. In order to answer research question three of this study (pertaining to the identification of specific attributes relevant to the assessment of school dropout dimensions using the identified datasets) individual questions or data variables are identified per sub-theme which provides some insight in describing how learners or the surrounding community (if relevant) perform as per the results of the collected data.

Furthermore, as each dataset has been assessed, the data quality score of the dataset can be applied to the particular question to assess how reliable the data used to describe the sub-theme actually is. In Table 56 (see page 189) the SDCAF is applied in combination with the results from the PDQAF with a specific focus to the learner's physical well-being. The colour scheme presented in the table highlights data gaps (red cell), datasets with poor data quality with a score below -5 (red cell), datasets with questionable data quality that have attained a data quality score between -5 and 0 (pink cell), datasets that have an acceptable data quality score between 0 and 5 (light green cell) and lastly, datasets with a high data quality score which is above 5 (dark green cell).

In review of the available datasets per thematic area, the most data collection gaps that were found related to the learner's physical well-being was in the theme of menstruation and female maturation, with a lack of available data found across each of the three countries. Although South Africa has 8 identified datasets in comparison to Brazil's 2 and India's 3 for this particular theme, the bulk of data questions which are found, are sourced from datasets with poorer data quality scores. Relevant data related to this functioning from the higher scoring datasets such as the Census, Community Survey, NIDS or TIMSS, are not frequently found, thus indicating that the crucial and rare variables are captured in lower data quality data collections.

Theme	Sub-Theme	Brazilian Data Sources	Indian Data Sources	South Africa Data Sources
Malnutrition and Hunger	Access to feeding programmes	**PNDS 2006** • They receive basic food basket inclusive of dairy, vegetables, and other food stuffs	**DISE 2014/15** • Schools Providing Mid-Day Meal (Government & Aided Schools) • Schools where Mid-Day Meal is Provided and Prepared in School Premises (Government & Aided Schools)	🟥
	Number of learners affected by malnutrition	**PNDS 2006** • Nutritional Index, height and weight for age appropriate	🟥	**TIMSS 2011** • Students suffering from lack of basic nutrition **YRBS 2011** • Percentage of high school learners who are undernourished and over-nourished by gender, race, grade, age and province **DHS 2003** • Nutritional status of children • Nutritional status of children at birth
	Nutrition intake of learners	**PNDS 2006** • Consumption of particular food stuffs over last 7 days • M481-Indicated dose of vitamin A, in the last 06 months	**DHS 2005/6** • Micronutrients, Vitamin A intake • Nutrition - child feeding practices, vitamin supplementation, anthropometry, anaemia, salt iodization	**DHS 2003** • Micronutrient intake among children
	The effect of extreme poverty and hunger	**PNDS 2006** • Adult skipped meals • Experiences of hunger, frequency of eating an insufficient amount of food • Food consumption aged below 18 years • Hunger aged below 18 years	🟥	**CS 2016** • Ran out of money to buy Food in past 12 months • Households who skipped a meal in the past 12 months, **SASAS 2012** • To what extent was the amount of food your household had over the past month less than adequate, just adequate or more than adequate for your household's needs?
Menstruation and Female Maturation	Gender biased communities	**PNDS 2006** • Married before age of 20	**Census 2011** • Ever Married and Currently Married Population by Age at Marriage • Duration of Marriage and Educational Level -2011 • Number of Women and ever Married Women by Present Age, Parity and Total Children Ever Born by Sex	**DHS 2003** • Age at first marriage • Median age at first marriage • Polygamy

Theme	Sub-Theme	Brazilian Data Sources	Indian Data Sources	South Africa Data Sources
		🟩	**DHS 2005/6** • Women's status • Women's Empowerment - gender attitudes, women's decision making power, education and employment of men vs. women • Gender/Domestic Violence - history of domestic violence, frequency and consequences of violence	**SASAS 2012** • Girls and boys should be educated separately
	Access to female sanitary products	🟥	🟥	🟥
	Understanding female maturation	**PNDS 2006** • Knowledge of times when at risk of falling pregnant	🟥	🟥
	Cost of female sanitary products	🟥	🟥	🟥
	Access to adequate sanitation in schools	**School Census 2015** • Type of sanitation	🟥	🟥
	Provision of medication to counter menstruation discomfort	🟥	🟥	🟥
Teenage Pregnancy	**Provision of sex education**	**School Census 2015** • Training specific to Gender and Sexual Diversity	🟥	🟥
	Pregnant <20 years	**PNDS 2006** • Number pregnant below 20 • Reasons for falling pregnant, if below 20	**DHS 2005/6** • Desired family size • Fertility and Fertility Preferences - total fertility rate • Marriage and sexual activity	**Census 2011** • Age at first Birth **CS 2016** • Children ever born **NIDS** • Currently Pregnant? **DHS 2003** • Teenage pregnancy and motherhood **DBE ASS** • Number of learners in ordinary schools who fell pregnant, by province and grade

Theme	Sub-Theme	Brazilian Data Sources	Indian Data Sources	South Africa Data Sources
	Accessibility of contraception	**PNDS 2006** • Knowledge of various forms of contraception	**DHS 2005/6** • Family Planning • Knowledge and use of contraceptives	**YRBS 2011** • Percentage of high school learners who used various methods of contraception by gender, race, grade, age and province • Percentage of high school learners who always use condoms, who had either been pregnant or made someone pregnant, and who has a child
				DHS 2003 • Knowledge of contraceptive methods • Condom use at first sex among young women and men • Current use of contraception • Current use of contraception by background characteristics • Current use of contraception by women's status • Ever use of contraception • Higher risk sex and condom use at last higher risk sex in the last year among • Multiple sex partnerships among young women and men • Number of children at first use of contraception • Sexual activity and condom use in last 12 months • Source of contraception • Timing of sterilization
	Late School Entry	**School Census 2015** • Distribution of Leaners by Grade and Age	**DISE 2014/15** • Distribution of Enrolment by Age & Grade: All Areas	**DBE ASS** • Number of learners, by age and grade

Table 56: School Dropout Capability Assessment Combined with the Public Data Quality Assessment for learner's physical well-being

8.2.1. Malnutrition and Hunger

In terms of malnutrition and hunger, 4 sub-themes are identified, viz., the learner's access to feeding programmes, learners affected by malnutrition, the nutrition intake of learners and lastly the effect of extreme hunger and poverty. As discussed in section 3.4.3, the malnourished learners tend to be more prone to disease and are predisposed to cognitive underdevelopment. This factor combined with others, tends to pull learners out of schools. Each of the countries have school feeding programmes, which are not reported by the national education departments in Brazil or South Africa. In this area, no data on the school feeding programmes is available, whilst in India the Unified District Information System on Education provides data on the number of schools providing a mid-day meal as well as the schools that prepare such meals. In the absence of alternative data sources with a higher data quality, the U-DISE data is recommended. In Brazil, the Demographic and Health Survey provides the prevalence of access to food baskets which is not comparable to the data from India. Therefore, in Brazil, there is also a gap in terms of data pertaining to school feeding programmes. However, the data from PNDS 2006 is of a high data quality level and is recommended for reporting.

In terms of the number of learners affected by malnutrition, data from India is missing whilst Brazil and South Africa measure comparable data items. In Brazil, a nutritional index is provided together with an aggregation of the body mass index across the surveyed respondents. This differs from South Africa’s TIMSS which provides the prevalence of students suffering from a lack of basic nutrition. This is comparable to the data from the YRBS which provides the percentage of high school learners who are undernourished and over-nourished. In this regard, due to the similarity of the data questions, the TIMSS data is preferred over the YRBS due to its higher data quality score. Data from the DHS of 2003 is also available. It provides the nutritional status of children (children at birth in particular) but the data is more outdated than the TIMSS information. Since the DHS has not adequately defined its data fields such as nutritional status one recommends the use of the TIMSS data.

With respect to the nutrition intake of learners, data is available from all three countries. South Africa’s data from the DHS has performed poorly in terms of data definitions, whilst India’s and Brazil’s data is available from high data quality sources. In South Africa the DHS reports on the micronutrient intake of children in general, whilst India reported on data across various factors such as Micronutrient Vitamin A intake specifically and general nutrition practices such as child feeding, vitamin supplementation, anthropometry, anaemia and salt iodization. The questions posed here focused on the consumption of such nutrients in the previous day whilst in Brazil, the questions focused on the consumption of nutrients over the past 7 days. The specific concerns within the PNDS is the consumption of particular food stuffs over the last 7 days and whether the respondent received the indicated dose of vitamin A during the last 6 months prior to completing the survey. Although, each country collects information on the nutrition intake, one must carefully consider the definition of the variables concerned. The data collected from the three countries in this instance discuss similar yet incomparable measures. For example, the Vitamin A dosage related questions between Brazil and India differed as the PNDS counted those that received a Vitamin A supplement in the last 6 months, whilst the Indian collection referred to the number of infants that received Vitamin A in their first 2 months after birth.

With respect to the effect of extreme hunger and poverty, the sub-theme is quite broad in this respect. However, none of the examined datasets provided any questions or variables which discussed these concerns in India. In Brazil, the PNDS of 2006 was again found relevant, whilst in South Africa there were 2 surveys, viz., the Community Survey of 2016 and the South African Social Attitudes Survey (SASAS) of 2012. As noted, the PNDS was found to be of high data quality, and South Africa's Community Survey was found to perform strongly in terms of data quality. However, as the detailed community level information is not yet released, the level of analysis possible at this stage is limited till the dataset is published. The SASAS performed positively but with various data quality cautions that were noted. In terms of the questions that were posed to the survey respondents, the PNDS counted the number of adults who skipped meals which differed significantly from the Community Survey which posed the questions to the household and not the individual respondent. This highlights the care that must be taken not to mix data collections of households and persons. The SASAS data collection differs from the other 2 surveys as the SASAS measures the assessment of adequacy of the access to food and the number of respondents who skip meals. The adequacy of the amount of food is comparable to Brazil's PNDS questions related to the frequency of eating an insufficient quantity of food.

8.2.2. Menstruation and Female Maturation

The theme of menstruation and the biological changes that females undergo as part of maturation is found to have a direct bearing on the learner absenteeism at first and learner dropout over time. As discussed by Jewitt and Ryley in section 3.4.3, the school's ability to cater to these concerns helps in promoting gender equality and leads to higher female enrolment. However, despite the importance of the theme, the available data related to the theme is scarce. Perhaps a more exhaustive scan of surveys in the three countries would surface additional instruments. However, at this stage, the national statistical agencies and ministries of education in particular are urged to collect data in respect of the sub-themes such as access to female sanitary products, the learners understanding of female maturation, the access to adequate sanitation in schools and the provision of medication to counter menstruation discomfort.

Although, some data has been collected in terms of the gender bias in communities, the available data relates primarily to marriage of learners under the age of 20. On the positive side, the collection of such information is comparable as the number of female learners married at a particular age can be aggregated to a similar category when individual ages are requested as part of the questionnaire. This is the case with the data from the PNDS, the Indian Census and the South African DHS survey of 2003. Whilst the Brazilian questionnaire poses the question in terms of all learners below the age of 20, the Indian and South African data can be aggregated to follow this particular definition for comparative reporting purposes. However, apart from marriage related questions, the gender bias of communities is a difficult sub-theme to assess. The Indian Demographic and Health Survey (which was assessed strongly in the terms of data quality) attempts to furnish new data regarding the gender dynamic in India with questions on women empowerment and gender violence but such information was not identified in Brazil or India.

In terms of the data gaps under this particular theme we find no data available in relation to the access to female sanitary products, the cost of female sanitary products and the provision of medication to counter menstruation discomfort from any of the three countries. Data related to these sub-themes would be valuable to assess the needs of communities and provides a more explicit exploration of the inequalities of gender with respect to the greater costs that female learners and their families incur, which can be more restrictive to female learners. In terms of understanding female maturation and the access to adequate sanitation in schools, some data is collected by the PNDS and the School Survey in Brazil, with no similar data collections in either India or South Africa. With no alternative data source to the School Survey or the PNDS, both surveys are recommended for data use.

Considering the data quality of the surveys that are relevant to Menstruation and Female Maturation, the PNDS of Brazil and India’s DHS were of the highest quality and are supplemented by the School Census in India and the DHS and SASAS in South Africa. Each of the surveys presents relevant information pertaining to different aspects of gender bias. The questions that are identified could be found relevant in the other countries, therefore data collectors from BRICS should pay attention to such data gaps especially when they face similar challenges.

8.2.3. Teenage Pregnancy

In unpacking teenage pregnancy, four sub-themes were identified as relevant, viz., the provision of sex education to learners, the number of learners pregnant before the age of 20, the accessibility of contraception and learners that enter the school system late and become more inclined to start a family whilst attending school. As discussed in section 3.4.3 by Gustafsson, Hunt and Soares, teenage pregnancy affects learners universally across the three countries, as the pregnant female learner's attention is diverted entirely away from school attendance to their immediate concerns of their livelihood and that of their future child's. Therefore, the sub-themes that are outlined tests the learner's understanding of how pregnancy occurs and their options to avoid such circumstances. It was found that data regarding the age at when pregnancy occurs, the use of contraception as well as late school entry is available across the three countries. The provision of sex education is limited only to Brazil.

With respect to data on learners who fall pregnant below the age of 20, data is comparable between Brazil's PNDS and South Africa's Census of 2011. Both surveys are of high data quality and provide details pertaining to the age of the pregnant learner. In order to report on the data, the South African data on single age and pregnancy would need to be aggregated to be reported together as the PNDS reports on learners pregnant before they reach the age of 20. In India the data collection is more nuanced and the public's attitudes to pregnancy by the age of the respondent are available for cross-tabulation from India's DHS. However, the survey does not specifically ask the question of the age of a person's first birth. In South Africa, additional data is available from the Community Survey, National Income Dynamics Study (NIDS), DHS and the Annual School Survey which ask questions about pregnancy without requesting the learner's age. Due to the ranges in data quality, the data from the Census, Community Survey and the NIDS is preferred. If the data user requires trend data, the Annual School Survey, despite its poor data quality assessment, provides a time-series of such data.

The accessibility of contraception provided a varied range of data questions especially within South Africa. The Youth Risk Behaviour Survey (YRBS) and the DHS provide a series of questions which refer the use of contraception and experiences of sexual activity. The use of contraception is comparable between surveys in South Africa although the surveys are separated by 8 years. The question pertaining to the learner's knowledge of the various forms of available contraception is comparable between the PNDS, India's DHS and South Africa's DHS. Although the data quality from the South Africa carries various cautions, comparable data across the three countries are rare and the resultant data would present interesting results despite the time difference in data surveys.

Lastly, the final sub-theme related to the teenage pregnancy is the leaner's late entry into the school system. If the learner starts school at a late age, the learner is more likely to exit puberty whilst still in school and possibly consider starting a family whilst attending school. Fortunately, all three countries collect data on the learner's age and grade whilst in the school system thus allowing an analysis of the distribution of learner's older than the norm per grade. Data is in this regard is available from datasets of 'acceptable' data quality viz., the School Census in Brazil, the Unified District Information System in Education in India and the poorly performing Annual School Survey of South Africa.

The strong cautions attached to the Annual School Survey of South Africa results in a data user reluctantly using such data to make a comparison against high and medium quality data from Brazil and India. However, in the absence of alternative options, one is forced to use what is available. In this particular theme, various comparisons can be made in terms of a learner's knowledge of contraception as well as the distribution of inappropriately aged learners in the school system. When comparing, care must be made to ensure the particular options of the responses of data questions must be checked for consistency. For example, if a questionnaire suggested abstinence from sexual activity as an option for contraception, such an option would be incomparable against female condoms at a granular level. However, an argument could be made that the learner has some awareness and understanding of contraception as a concept despite the differing response options that are offered.

8.3. Data visualisation of comparable trends

In order to convey examples of some of the possible analysis using the datasets referred to above, the graphs provided at the end of this section have been prepared, which aggregates some of the comparable data referred to in Table 56. The arrows included within the table highlight which specific questions within surveys are presented to the user in a similar manner. The graphs that have been presented in Figure 8 to Figure 11 highlight that despite differing methodologies, the context of the questions offer enough similarities to help describe common trends in the selected countries, where similar types of questions are asked. In providing such analysis, care needs to be taken to present the different data sources clearly. Such graphs help illustrate the effects of different policy positions amongst the countries. The data presented is limited to only a few data surveys using the latest released results. Where previous iterations of the survey are available there are opportunities to present a time series of such variables in the chosen analysis to better depict effects of policy on the education system over time.

When attempting to compare data from different surveys, as in the examples provided, the data user should be aware that there are certain limitations to the types of comparisons that can be made. Effort must be made to present items of a similar context together. This requirement covers multiple concerns including the geographical context, time frame, the unit of measure, the targeted respondents and the particular type of measurement to which the survey is geared to produce.

In terms of the geographical context, data that is aggregated to a particular geographical level should be presented against data from other surveys at a similar level, for example, data from Brazil aggregated to the national level in the PNDS 2006 survey should be presented against other nationally aggregated data from the other survey. If the data is aggregated to a lower geographical level and presented in the same graph, care should be made to express the rationale for highlighting the specific geographical change.

With respect to varied time frames in a common graph, the data user should explicitly identify which period to which the data refers. Furthermore, data describing the trend should be aggregated to common scale and therefore, for example, comparing data of a month against a year should be avoided. In the examples provided below, we find the various surveys identified all relate to different years. Therefore, care was taken to explicitly identify the year to which the data variable refers.

The next important concern to be considered when comparing diverse data surveys is to refer to a common statistical unit of measure. For example, the number of observations related to a particular variable in a survey should rather be presented as a percentage instead of a physical count if presented against data from another survey, as the sample sizes differ resulting in a nonsensical comparison. Furthermore, when making a comparison, it is important to ensure that the targeted respondents of a question are compared against a similar group from another survey. For example, if survey X presents women’s responses to the use of contraception, the comparative survey should also present the responses of women, ensuring that responses of a single gender are compared against another.

The last key concern in terms of comparing data across surveys is to ensure that similar data points are compared amongst surveys. Physical prevalence rates should not be compared against perceptions. The data user should consider the type of data that is collected in a survey and be aware if the data measure refers to an attitudinal position adopted by the respondent or actually involves a physical measure. Using the data highlighted in relation to physical well-being, both types of surveys have been identified. Some surveys also collect both types of measures. The survey could count the number of occurrences of a particular concern and also collect information on the respondents feeling towards the occurrence, such as commonly found in the South African Social Attitudes Survey which measures attitudes and physical occurrences within the same survey.

Considering the above concerns when comparing data, within the examples presented, one finds that data from the variety of questionnaires can be presented together in a single graph. In producing these graphs, the country, survey source and time frame must be clearly presented to denote where the data comes from and what context it represents. Furthermore, attitudes are only presented against other attitudes and all measures are presented in terms of percentages to ensure comparability across the surveys with all surveys represented at a national level. Lastly, the specific questions of the survey are provided to ensure that the analysis clearly identifies the context to which the measurable dataset refers.

From the four graphs that are presented, a few interesting observations arise. In terms of the experiences of hunger, similar data questions were found in Brazil’s PNDS of 2006 in comparison to South Africa’s SASAS of 2012. Whilst the datasets are 6 years apart, the focus on hunger is asked in a different yet similar manner across the two surveys. In Brazil, hunger is referred to in a number of ways. Three of the questions asked are presented below in Figure 8. Each of the questions relate to the respondent’s experience of hunger in the past 3 months which differs from the SASAS question which refers to the adequacy of food in the household over the past month. Therefore, although a longer reference period was required in Brazil, the differences are expressed per reported variable for the purpose of clarity. Using percentages, and as the samples were both nationally representative, one can elicit from the graph that 26% of South African respondents in 2012 felt their access to food in their household was less than adequate in comparison to 7.76% of Brazilian respondents in 2006 who felt hungry over the course of the previous 3 months. Thus, whilst the variables differ, they present to the reader a common contextual message about hunger, which can thereafter be explored in greater detail.

In reference to the provided graph on the age of survey respondents when they were married, one notes that although the question amongst the three surveys of Brazil, India and South Africa are similar, the non-specific question on age has led to a limitation in the analysis that can be performed. In Brazil, the question is focussed on the number of respondents who are married before the age of 20 without specifically asking what age a respondent was married at. In India, 2 age groups are referred to in the Indian Census of 2011, whilst in South Africa the question is focussed only on the age group of 15-19. If the three countries phrased the question in terms of the specific age, the data would be more comparable. Furthermore, if one looks closer at the results of Figure 9, one notes the issue pertaining to poor data quality in the reported results as the Brazilian results are skewed by an inapplicable data value of <Null> which is included in the results in relation to the age as when the responded was married. The value of <Null> cannot be confused with 'Do not know' or 'No reply,' but rather refers to a data capturing error which is expressed in the reported data. The question of teenage pregnancy relates to a similar concern about age. In Figure 10, we find South Africa's Census reported on the age of the respondents' first pregnancy which differed from the manner in which the question was presented in Brazil. In Brazil, the question is asked in terms of whether the pregnancy was 'undesirable' at the time, when the respondent was between the ages of 15 and 19. The notion of desirability of the pregnancy in this regard, highlights the challenges involved in comparing data across surveys. Whilst the Census of South Africa asks the question broadly, in Brazil, the interpretation of desirability would influence the responses to the question, thus reducing the comparability of the data.

In terms of the last graph presented in Figure 14, one finds how the difference in purpose of asking a survey question limits the comparability of data across countries. In Brazil, in the PNDS, the literature refers to the universality of knowledge of contraceptives such as condoms, therefore the question that is asked is more pointed towards assessing the respondents' understanding of the purpose of the contraceptive instead of asking the less sophisticated question of whether the respondent is aware of contraceptive methods as asked in India and South Africa. An interesting observation of the data available in India and South Africa is that they both result from local versions of a Demographic and Health Survey. In South Africa, the DHS metadata does not explicitly refer to the international DHS Programme and the similarities in questions or methodology, but on closer examination of the data and the manner in which it is presented, the questions are identically presented to the respondent. One can infer that South Africa's DHS follows a similar approach to India, although the metadata does not explicitly make such a reference. In this respect, these two particular surveys likely share more commonalities if closely compared.

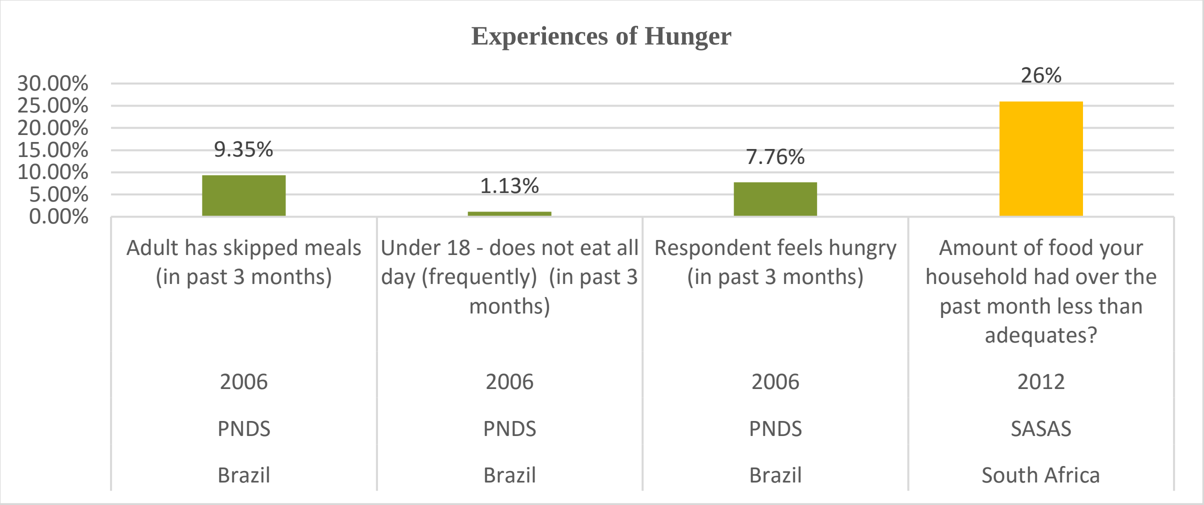


Figure 8: Data on Experiences of Hunger in Brazil and South Africa using PNDS 2006 and SASAS 2012

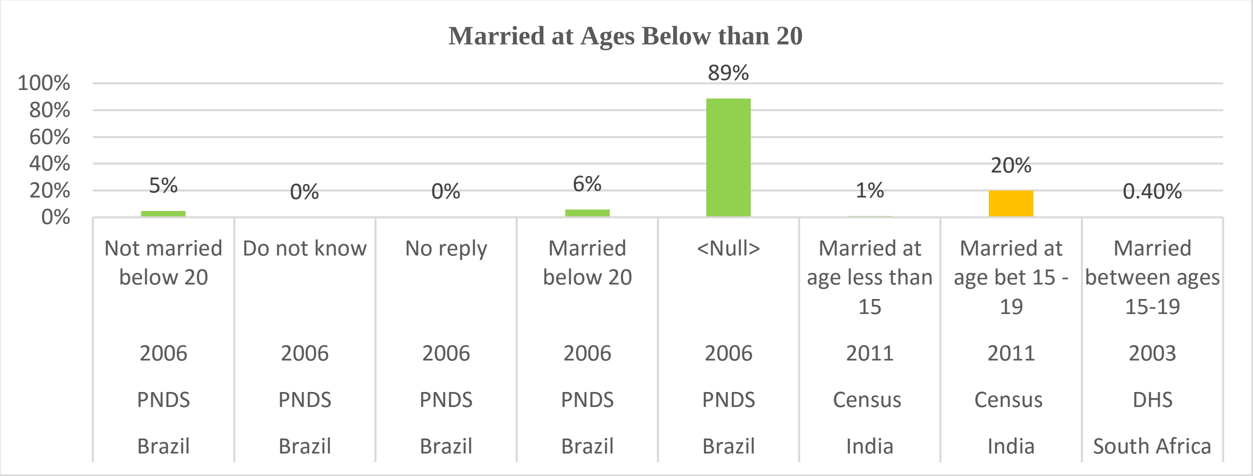


Figure 9: Data on Marriage before the age of 20 in Brazil, India and South Africa using PNDS 2006, Census of India 2011 and South African DHS 2003 (NB: <Null> refers to a data capturing error as reported in the PNDS)

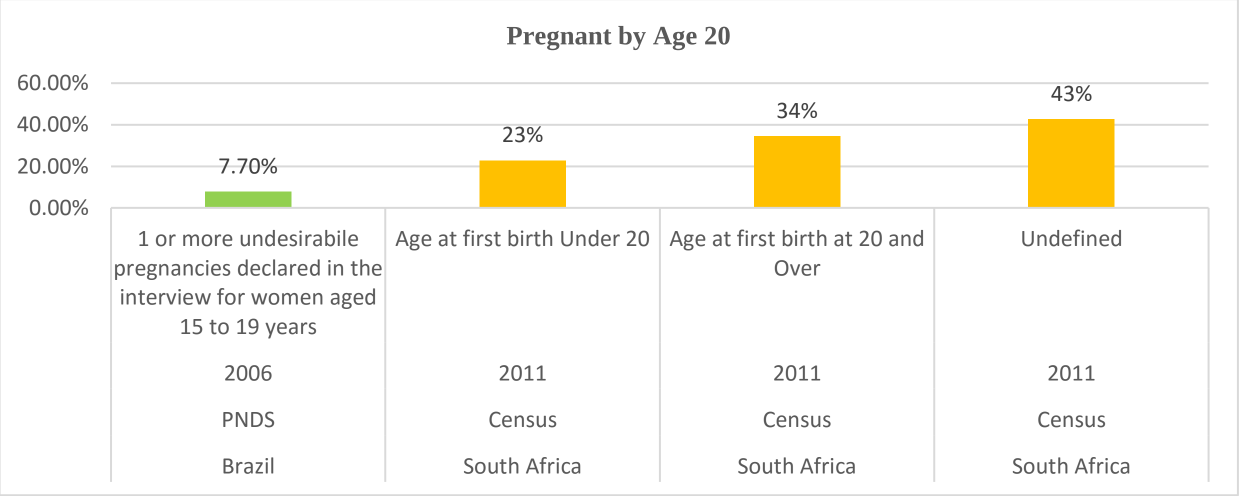


Figure 10: Data on Pregnancy before the age of 20 in Brazil and South Africa using PNDS 2006 and Census of South Africa 2011

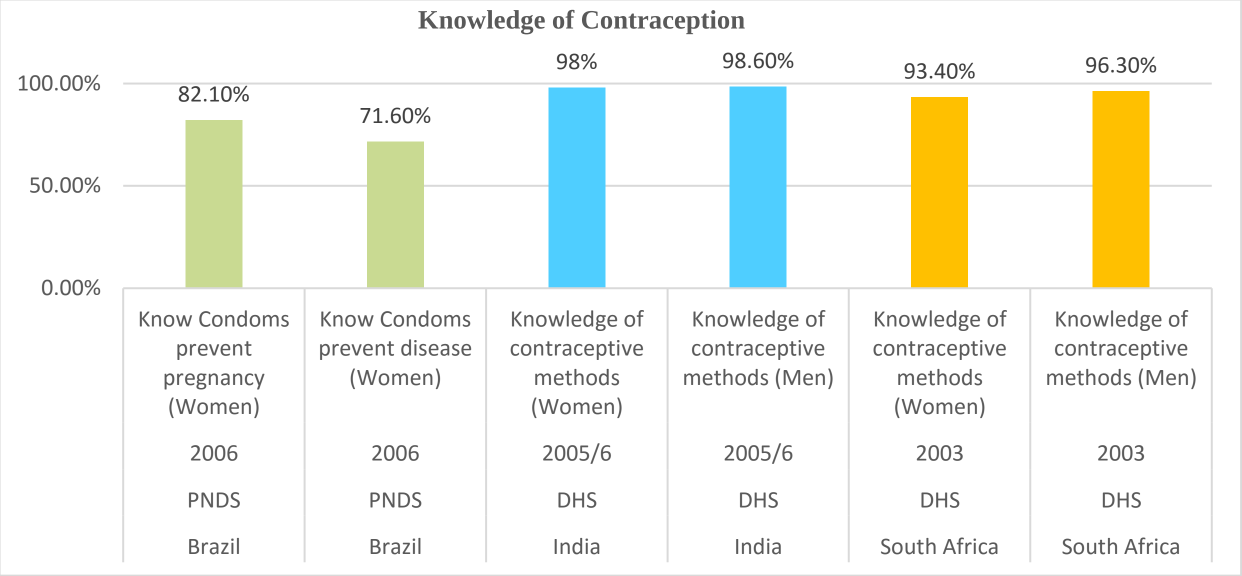


Figure 11: Data on the Population's knowledge of contraception in Brazil, India and South Africa using PNDS 2006, India's DHS 2005/6 and South Africa's DHS 2003

8.4. Does the practical data adequately describe the conceptual problem with respect to Physical Well-being?

In assessing the available surveys that were selected to provide data about the themes of general physical well-being of learners, one notes that questions posed are not often comparable. Questions that are phrased in surveys by the ministries of education, the statistical agents or other research bodies in the countries may cover common thematic areas but are often focused on the specific national need. Despite the gaps in certain countries for specific sub-themes (even when data is not comparable) some available data does provide an assessment specific to the country.

Comim identifies 4 conceptual concerns that need to be balanced against 3 practical needs as discussed in section 3.4.3. In assessing these factors one notes that Sen argued that the conceptual needs should always trump the practical considerations. However, to asses these factors, one must consider if the data provided presents suitable traits that express whether the valuational foundation suitably covers the sub-themes' core purpose, whether the broad ambit of human diversity needs are covered by reporting across the different demographics of the countries, whether the data objectively presents the sub-theme's purpose and lastly be aware that the data could present trends that are counterfactual to the argument that is expected when using the data.

When one considers the valuational foundation related to each of the sub-themes the major concerns that emerge with respect to the identified data primarily relates to the sub-theme of Gender Biased Communities and the Understanding of Female Maturation (apart from the themes where no data was identified). For these two specific sub-themes, the available data only captures a small-component of the nature of the particular area. For example, knowledge about the risks of pregnancy does not capture all the details related to the broad considerations of female maturation. From the human diversity perspective, a key need that must be met in the data is the granularity in terms of geography, gender, race, caste and creed where relevant to the country. In Brazil, the data within the PNDS was nationally representative but did not provide a field on population group whilst the School Census provided greater details regarding race in Brazil. In India, the Census data provided various extracts in terms of scheduled castes and religions. However, the data is provided in the form of static excel files, which limits the user's ability to produce cross tabulations of the many fields in the survey with these particular fields dynamically. The DHS data was more dynamic as the data was released in SPSS format with these fields easily accessible. In South Africa, the various surveys that were selected paid close attention to demographics and fields such as gender, population group and age provided in each dataset.

Importantly, the data provided is representative across each of the identified surveys where the fields are provided. In terms of objectivity, the fields that were identified across the various data collection instruments of the three countries were almost entirely objective in nature. The only subjective question that emerged was captured within the South African Social Attitudes Survey where the question that was selected referred to the respondent's opinion on the extent to which their food supplies were adequate as compared to an objective measure of hunger. All other questions focused on absolute observations instead of opinions of experiences. Lastly in terms of Comim's conceptual requirements, in terms of the counterfactual nature of data (from the 4 examples presented in the previous section) the data about the knowledge of contraception may be unexpected. This is because the knowledge of contraception is high but the impact of not using contraception is limited if one considers the high rates of unplanned pregnancy, specifically in South Africa. This highlights that knowledge of a possible solution to a problem does not guarantee the successful resolution of the problem.

From the data perspective of Comim's concerns, the data that is provided across the surveys has been granular and sufficiently detailed. The datasets are made available in wide tables in a multitude of data formats. The concerns that have emerged are from the Indian Census and South Africa's Annual School Survey where detailed information is not provided and the data is also not fully published or accessible. Lastly, in terms of weighting, the surveys such as South Africa's Census and Brazil's PNDS seem to have paid attention to the necessary calculations that assess the weighting of the data. However, the data with poor data quality, such as the Annual School Survey in South Africa and the Demographic and Health Survey in South Africa, have not discussed the weighting of the data in their released metadata.

There are some concerns that have emerged in assessing the performance of the data in terms of the conceptual requirements, but the primary concern of reporting on the conceptual requirements relates not to the data that is available but rather to the numerous data gaps which exist. Public data providers from the 3 countries should assess how best to prepare new data collection instruments which report on such sub-themes. In terms of the data gaps which have emerged the following situation exists:

a) With respect to the sub-themes of menstruation and female maturation, there are many gaps found across the three countries. Whilst there is some comparable data on early marriage in the three countries, there are various other issues which are not found to be available within the identified survey pertaining to gender biased communities. Whilst the DHS of 2005/6 presents some data on women empowerment and gender violence, no similar data was found in Brazil or South Africa. It may be that that gender related questions are not managed by the ministries of education or the statistical bodies that were reviewed. Based on the assessment of available data, Gender Biased Communities are assessed in a very limited fashion. Furthermore, the other sub-themes of menstruation and female maturation are very sparsely populated in terms of related data questions per thematic area. This signifies that education ministries in particular have not recognised how crucial issues such as access to and costs of female sanitary products or the provision of suitable sanitation for female learners in Brazil, India and South Africa are.

b) With respect to the Teenage Pregnancy thematic area, data pertaining to the provision of sex education in India and South Africa is limited and is a clear gap. Furthermore, in review of the type of questions that are asked in the 3 countries in relation to teenage pregnancy, there is a difference in approach between Brazil and the other 2 countries. Whilst India and South Africa seek to measure the occurrences of pregnancy, the PNDS in Brazil questions the reasons behind the pregnancy. Therefore, the depth in detail elicited in the Brazilian questionnaire provides greater substance than in India or South Africa. Furthermore, as the PNDS of Brazil is based on the DHS Programme backbone, India clearly follows the DHS Programme structure and South Africa seems to follow the DHS Programme without explicitly stating so in their metadata. There are great opportunities in the health space to build greater cohesion in the survey questions that are asked in future iterations of these surveys.

c) With respect to the effects of malnutrition and hunger, there is a great diversity in questions that are asked in relation to the identified sub-themes. However, there are few sub-themes that are comparable across the three countries. There are limitations regarding the specificity of terms used which negate possible comparisons especially in terms of nutritional intake of learners. The lack of comparable data is not necessarily a limitation to assess the severity of the challenges faced in each country. The concerns that do emerge are: the gaps available in South Africa regarding school feeding programmes; in India pertaining to the numbers affected by malnutrition; and the number affected by extreme poverty and hunger. The corresponding available questions posed in the other 2 countries are useful in identifying additional questions that could be included in the current questionnaires that are in operation.

Of the 13 applicable data collection instruments across the 3 countries only 2 surveys from South Africa were negatively assessed. However, various cautions were offered against the 5 surveys that did not perform strongly with data quality scores between 0 and 1.6. The effect of these weak scores impacts the value of the reported data of South Africa especially in terms of the menstruation and female maturation theme as well as the teenage pregnancy theme. With respect to teenage pregnancy, the Census, Community Survey and NIDS survey are all assessed strongly in terms of data quality. However, this strength is limited as the three surveys all report on a similar question on the number of pregnancies of teenagers with little additional insights offered.

In Brazil and India, the PNDS of 2006 and the DHS of 2005/6 provide a wide range of questions which support a greater number of sub-themes thus providing a stronger quality of data especially in relation to Malnutrition and Hunger as well as Teenage pregnancy. However, all three countries perform poorly in terms of capturing data on menstruation and female maturation and one needs to examine the reasons for this gap in data collection.

An additional benefit of performing this type of analysis (apart from identifying gaps in data collection) is that it helps to identify at a national level what types of insights can be learnt from BRICS partners. Whilst South Africa is producing some very high quality datasets in the Census, Community Survey and National Income and Dynamics Study, there are nuances in data collection which can be learnt from the partners. With greater cooperation between the countries South Africa could benefit from the insights gained in Brazil and India in terms of performing large and specialised data collections. South Africa could also assist their BRICS partners by sharing the advances they have made in developing the SASQAF. This could strengthen their data collection initiatives. Furthermore, whilst cooperation across countries would be beneficial to improve the quality of surveys, the intra-country cooperation would also be beneficial. Ministries and other research bodies that conduct surveys could also learn from the national statistical bodies when implementing quality assurance procedures. Another key factor which also impedes the assessment of data quality is the lack of detail within the released metadata from the education ministries and research bodies. Whilst these organisations may follow detailed data quality protective practices, their released metadata does not furnish the data user with sufficient details. In the absence of such information, the data user is forced to assess such data collection instruments poorly.

Chapter 9

Conclusion

9.1. Using public data produced by the BRICS for reporting on social indicators

From this investigation of the School Dropout Phenomenon affecting Brazil, India and South Africa, it is clear that these countries have expertise in producing data of a high data quality standard. The importance of using data in monitoring and evaluating policy implementation has been recognised across the BRICS in the manner that national statistics agencies are producing a greater number of data collection instruments that focus on the various challenges in their countries. Greater emphasis is also placed on ensuring that surveys produce results that can be trusted by data users, with a greater focus assigned to data quality protection. Today, data collectors are more connected than previously, thus leading to numerous international collaborations on data collection. With such collaboration, greater lessons on international best practices are shared more often. In this narrow field of learners who drop out of school, surveys such as the Demographic and Health Survey and the Trends in Mathematics and Science Study follow similar standards and guidelines, across countries, in the manner that questionnaires are rolled out and in terms of the data quality controls that are implemented. In time, a greater number of countries may partner in such studies due to the benefits of following a well detailed and planned survey collection process.

In addition to the surveys carried out by national statistics agencies, independent research organisations are also producing high quality data collection instruments such as the National Income Dynamics Study in South Africa, which collect data related to the particular priorities of their countries but also differing in their data collection approach to contrast their findings against other surveys that exist and to test the perceived assumptions about social trends. For example, the panel data approach employed by SALDRU in the NIDS survey supports a deeper form of statistical analysis in terms of modelling the complexities of human behaviour. A trend that is noticed across the 3 countries pertains to how representative such surveys produced by research institutions other than the national statistical bodies are. For example, those conducted by South Africa's Medical Research Centre or India's Demographic and Health Survey provides data representative only at the national level. These limitations stem from the identified sample size which focusses on producing data that describes the national trends. However, if one considers how the effects of inequality of the BRICS are experienced at a community level, broader samples are required. It requires organisations to invest to a greater extent in the rollout of such surveys to enable this deeper form of analysis.

From the application of the PDQAF, noted from Figure 7 and Figure 12, many of the datasets that were examined in this study were positively assessed. When the Statistics South Africa scale of statistics quality is applied to the examined datasets, in relation to learner's valuing their physical well-being, it is noted that 6 out of the 13 surveys were assessed as 'quality datasets,' and another 5 datasets were identified to present 'acceptable' data quality. This brief review highlights the options that are available when analysing a particular social phenomenon. However, whilst there are options to use to datasets of a high data quality, the usefulness of the data in individual scenarios must be ascertained. As identified in section 8.2 and in Table 56, data quality should be assessed from the context of what data is available for a particular theme or specific sub-theme which describes the broad aspirations of the learners affected in this study. In relation to the learner's valuation of being physically well, data providers in the countries (in the least, national statistics providers and ministries of education) have not recognised all of the valued elements that describe the social challenges that are prevalent, for example, as found in the theme 'female maturation' as noted in section 8.2.2. Whilst understanding that such social concerns are numerous and cannot all be captured in a survey, greater effort should be made by governments to assess what is valued by the target audience of their policies and thereafter institute quality data collection processes to gather relevant information. The gaps that relate to female maturation and teenage pregnancy, highlight the severity of concerns which impedes the capabilities of an entire gender in this instance. This emphasises the significance of conducting such an assessment.

On examination of the individual data collection instruments, the strongest performing datasets that were examined were found in South Africa and were managed by Statistics South Africa. This organisation manages South Africa's Census, Community Survey, General Household Survey and various other surveys not examined in this study. The strength of the Statistics South Africa processes follows from their adoption of the SASQAF which is embodied into all the data collection and dissemination processes followed by the organisation. Their processes provide a great example to the other BRICS countries regarding how to centralise data quality in all data processes that are followed. Both Brazil and India performed strongly in the rollout of their versions of the Demographic and Health surveys. The adoption and application of international standards in the metadata reports that were published drove the strong data quality scores that were achieved. On the opposite end of the data quality scale, the poorest quality datasets were also found in South Africa, especially all the data that is produced by the National Department of Basic Education (DBE). The complete lack of published metadata provides the data user no surety in terms of the data quality of the data outputs as no context is offered regarding the processes that were followed. Furthermore, DBE's data quality problems are compounded by their data dissemination strategy in which all data outputs are accessible via request alone. Some second-hand analysis is offered in publications such 'Education Statistics at a Glance,' but the detail offered is limited when one considers that the surveys target all the schools within the country. Recognising that there is great value in the depth of information collected by DBE, cooperation in introducing data quality collection standards as followed by Statistics South Africa would greatly assist to improve the department's data quality score. Thus, intra-country collaboration is recommended.

The datasets that were assessed more strongly in terms of data quality, shared a single common approach. The detail published in the accompanying metadata to a released dataset was of a high level of quality which helps to assure the data user in terms of the various dimensions of quality which they value. Where there are gaps in the metadata, the assessment of data quality in other dimensions such as clarity, accuracy and comprehensiveness suffers because the data user is provided no means to make an assessment. In the absence of published metadata, the data user is forced to assume the worst and therefore these dimensions are assessed poorly. The reality of the situation may differ from the data quality assessment but without any published information, the negative assessment must stand. Considering the requirements of the data user, it is better for organisations to make an effort to publish details about their data collection processes. If organisations such as the DBE or those responsible for the Crime Statistics of each country released some documentation the data user would have a better sense of how the data can be used. Furthermore, when assessing the policy implementations of the BRICS, having confidence in the reported statistics is crucial. Furthermore, exposing the challenges that are experienced may provide researchers' the opportunities to better understand how the data is collected and thereafter assist with recommendations to improve the process.

In answering the first research question of this study, it is found that the 3 countries do produce datasets of a sufficient level of data quality. However, this endorsement is tempered as certain data gaps have been recognised for specific sub-themes related to the school dropout phenomenon. Data collectors should heed the existence of these gaps and identify other areas that present similar challenges by perhaps following a similar process as outlined in this study.

9.2. Application of the School Dropout Capability Assessment Framework and the Public Data Quality Assessment Framework

The second research question of the study asked what possible framework could be adopted to unpack the challenges that Brazil, India and South Africa face in terms of school dropouts and data availability. In answering this question, two frameworks were identified: a capability assessment framework and a data quality assessment framework. The SDCAF is specific to the school dropout context and is structured on the shared challenges which affect the three countries and other developing to middle-income countries. It is not possible to apply the framework as-is to another social challenge affecting these countries or the wider BRICS as these issues are context specific. However, the broad outline of the framework and the steps followed are sufficiently generic to be adapted to a different scenario. Alongside this application, the PDQAF can be adopted in the analysis of any publicly released dataset across the BRICS as the data quality dimensions are not unique to Brazil, India and South Africa. This was found after the data quality priorities of the three countries were found not to be exclusive of issues of data quality.

On a closer examination of the SDCAF, the process of itemising the set of valued functionings into a Capability Set for the assessment of an individual's real freedom can be applied in all social contexts. The key requirement is identifying what the central real freedom is. This must be addressed by the particular framework. For the purposes of this study, the real freedom as identified in section 3.4.2 was the 'Learner's real freedom to complete school in preparation to access employment opportunities and an improved quality of life.' Whilst this freedom is a complex and subjective concern, the strength of the Capability Approach is that it emphasises detailing the plurality of valued states (functionings) which an individual aspires to attain. In this regard, the core task that must be performed is enumerating this core set of valued states. The process of enumeration has empirical qualities which promotes the assessment of a complex phenomenon quantitatively which leads to the incorporation of the approach in studies such as the Human Development Index. The identification of the array of valued states allows for the identification of related indicators or datasets which can be used to describe the phenomenon. It is noted, that quantitative data aggregation cannot describe the broad conceptual requirements of the individual functioning but, the strength of the approach is that it moves the analysis of human development factors away from commodity driven measures to a better set of factors which better describe the needs of the population with regard to challenges such as poverty and inequality. It is this process of defining real freedoms and thereafter valued functionings in the form of a Capability Set which allows one to assess the real freedom an individual has to attain those valued states to which it aspires.

Comim et al. (2008) outlined a sequence of tasks which can be followed in carrying out a capability assessment and it has been followed by the Human Development and Capability Association in the many applications of the approach that the organisation has conducted. These tasks are described in section 5.2 of the Methodology chapter and are provided to guide the analyst to ensure that the practical implementation of their study is bounded in terms of conceptual requirements of the particular phenomenon that is assessed. The challenge that was experienced in producing the SDCAF, is that generally the applications that Comim et al. referred to were of a smaller scale, referring to a single community that was analysed in terms of a yet to be defined questionnaire. In these instances, the structure and content of the questionnaire was determined based on the analysis of the environment that is surveyed. The difference in this study is that relevant questions from datasets are identified after the identification of the Capability Set. Therefore, it is possible to have data gaps when describing the sets of themes and sub-themes which describe the identified functionings. Hence, a critical step which this study advocates for also considers the criticisms made of the HDI and other Capability Approaches to test the data quality of the selected appropriate datasets.

In summary, whilst the SDCAF in its entirety is not reusable, the tasks outlined in the methodology of the study can be applied in different contexts to determine what is most valued by the populations of the BRICS. Alternatively, the PDQAF is completely reusable as the outline of data quality dimensions is relevant across Brazil, India and South Africa. The only impediment to its application in China and Russia is if these countries have expressed policy positions which are contrary to the tenets of quality data collections. This is unlikely considering the joint call for the identification of methodologies which are applicable to all BRICS countries.

9.3. Challenges to Data Quality in the BRICS

Although this study examines only a small subset of the available datasets from the BRICS which examine social trends, the challenges that were identified are likely to be shared in other data collection instruments and found in other social contexts. However, before the specific data concerns are listed, the broad challenges applicable to the BRICS must be discussed.

When sourcing data to describe a particular functioning included in the Capability Set, one must be aware the diverse nature of the BRICS which leads to data that is often context specific as the questions that are posed are based on the needs of the country. This diversity, limits the comparability of the questions in certain instances. For example, considering the data presented in Figure 11, one should note the subtle differences in the questions asked in Brazil, versus those asked in India and South Africa. Although the questions in the surveys probably follow similar data standards of the international Demographic and Health Survey Programme, the questions in Brazil are nuanced differently. In India and South Africa, the question only asks if the respondent is aware of a condom as an option for contraception, whilst in Brazil, the question is slightly more sophisticated and asks if the respondent is aware that a condom is used to prevent the spreading of disease or to prevent pregnancy. The included caveat in Brazil limits the comparability of the data questions. However, the inclusion of the caveat in this instance follows from a need in Brazil for a deeper form analysis.

Furthermore, an additional broad challenge facing the BRICS that was included in the BRICS Heads of State call for statistical cooperation amongst the countries, is the need to identify a joint set of standards and methodologies when producing data surveys. The end goal for such an initiative is to promote data comparability. However, since the call was made in 2014 the BRICS have not introduced any processes to adopt common standards. However, considering the magnitude of the task required to standardise all data collection instruments in the BRICS and, also considering that the surveys are tailored to the specific needs of the countries, achieving this goal will be a long term effort. In the meantime, the BRICS need to investigate what options are available for data reporting whilst the data questions do not entirely refer to the same concepts. Section 8.3 outlines a few pointers that the BRICS can follow when reporting on similar but different datasets. These involve reporting the source, time-frame, and specificity of the data question that is involved. In the long term, the BRICS will need to assess how feasible it is to adopt a common methodology across the numerous data surveys that are carried out. In addition, they should consider whether a selection of surveys can be made which can start to follow a common approach. As was found by the IMF, when countries adopted the common DQAF, a by-product of the framework was the that the countries began to share a common language when discussing data quality challenges. Therefore, an adoption of a common data quality framework, may spur the BRICS into better understanding how to tackle the data methodology challenges that they face.

Lastly, a significant challenge affecting the layperson in the BRICS and the ability to understand the data across countries, is language. In Brazil, the primary language used to report data in is the Brazilian variant of Portuguese, similarly though not discussed in this study, one notes that China's data is reported in Mandarin and Russia reports in Russian. The BRICS statistical office will need to identify a process to translate the data collected into languages of the BRICS countries to promote the sharing of information.

9.4. Findings and Assessment

Apart from the broad challenges which affect data quality and use of data, various findings were made pertaining to the individual datasets reviewed in terms of the physical well-being of the learners. In determining whether the conceptual needs of the assessment were met based on the use of the available data, it was found, using Comim et al.'s 4 factors, that the valuational foundation of the particular themes were adequately expressed. The concerns that emerged related to the gaps in data collection specifically in terms of gender biased communities, female maturation, access to and costs of female sanitary products, access to school sanitation services, the provision of medication to counter menstruation discomforts and the provision of sex education to the learners. The lack of data in some of the sub-themes is a primary concern. In addition, some of the data available in terms of gender biased communities and female maturation did not exhibit the necessary depth and complexity of the particular theme. It is recommended that the data providers unpack these sub-themes and identify suitable questions to quantify the related issues.

The other conceptual issue of Comim et al. that can be noted is the human diversity perspective where one needs to assess whether the datasets that were used support the analysis of the identified data questions in terms of the various marginalised population groups in each country. In this regard, the datasets seem to adequately include the under-represented communities in each of the countries. The only survey where this question was excluded was in Brazil's Demographic and Health Survey (the PNDS of 2006) where population group question was either not asked within the questionnaire or not provided in the reported dataset. Largely, in terms of this requirement, marginalised communities were represented.

With respect to the type of data quality concerns that emerged (which also answers research question four of this study, which requires the identification of impediments to data reporting), the datasets impacting the assessment of the learner’s physical well-being such as Brazil’s School Census, India’s District Information System for Education and in South Africa, the Youth Risk Behaviour Survey, the Demographic and Health Survey, the South African Social Attitudes Survey and the Annual School Survey were assessed negatively in terms of many data quality dimensions. The worst assessment was that of DBE’s Annual School Survey that was assessed poorly as the DBE did not publish their metadata or release their detailed data publicly. The other survey’s made some effort to publish their metadata but excluded pertinent information such as codebooks with concept definitions or supporting information such as response rates to assess data coverage, imputation rates to assist in the assessment of accuracy or excluded details about the data quality constraints which support the assessment of the data integrity dimension. Reasons for the exclusion of such supporting information needs to be better understood. If such processes are conducted and not documented the assessments would improve only if such details are included in the released metadata. For such occurrences, organisations would benefit from adopting a framework as the PDQAF or South Africa’s SASQAF if they are agreeable with the requirements for such reporting.

The intention of the study is not to identify comparable datasets but rather to explore the quality and availability of data pertaining to a particular theme. However, in some instances it was found that surveys which followed common standards were comparable although the time-frame of the surveys differed. It suggests that it is possible for surveys that are run independently in the BRICS to share a common approach. The guidelines provided in such surveys promote data quality as they denote specific data quality requirements that must be conducted. These are common to the data quality dimensions expressed in the PDQAF.

9.5. Recommendations

Over the course of the study, various concerns have emerged which has led to the following set of recommendations.

In assessing the data quality of the various datasets that were identified in relation to all the identified functionings of the Capability Set, a common challenge which was found across all data collection instruments in all countries relates to the manner in which one could measure data quality. None of the surveys make a reference to a chosen measure of data quality in their reported metadata. In producing the PDQAF a rudimentary measure was introduced to convey the aggregated assessment of quality across the identified data quality dimensions. However, a deeper exploration is needed of how one could quantify such a measure. Such an approach could be explored in an expansion of the PDQAF or an adaption of Statistics South Africa's SASQAF.

Furthermore, it is noted that the surveys that adopted the SASQAF performed the strongest in terms of the data quality assessment. This is due to the manner in which data quality is embodied in the data collection process of the data provider. It is therefore recommended that the BRICS adapt a suitable variation of the SASQAF for their data collection practices. In terms of possible variations that could be made to supplement the current dimensions of the SASQAF, it is recommended that the SASQAF also introduces a template for reporting metadata. This is to ensure that the organisation not only follow data quality processes internally but, also communicates these practices to the public in their released metadata.

In terms of condensing the assessment of school dropouts across the BRICS, the Capability Approach has a natural progression towards the adoption of a composite index informed by the selection of indicators for the enumeration of functionings within the Capability Set. However, considering the data gaps which exist within the assessment of being physically healthy, it is not recommended that a standardised BRICS wide composite index is created to describe the phenomenon. If one considers the variations in the questions that are asked within the surveys, it is more appropriate to produce a country specific composite index based on measures specific to the context of the country. These measures could be used to assess the same selection of functionings across the BRICS but, need not necessarily share the same set of indicators. Over time, with a standardisation of data collection processes, such an index could be created. Furthermore, based on the aggregated summation of measures related to the composite school dropout index, one could track progress using the country specific composite. In this light, analysts could report on growth or decline of the country specific indices. However, based on the incomparability of the index, it would not be possible to compare the results of the index across the BRICS until the BRICS can agree on a common set of indicators and collect data following a common approach.

In terms of intra-BRICS comparisons of similarly themed datasets, it is recommended that analysts follow the examples offered in section 8.3 in terms of how graphs are used to report on data questions which are presented similarly but possess certain differences in terms of how the question is posed to the respondent. Such graphs and tables should ensure the reported data clearly indicates the country, source, time-frame, geographic aggregation and present the data question clearly to the respondent.

The data quality scores that were attained across the surveys are largely based on the level of detail included within the published metadata. Firstly, it is recommended that the organisations make a concerted effort to separately report their metadata in a manner distinct from the findings report which was found with some data collections. When the metadata was included as an appendix to a report, it was generally found that the metadata lacked sufficient information pertaining to each of the data quality dimensions. Secondly, organisations that are not primarily responsible for data surveys, but conduct such data collections, require support in managing their data collection as well as the manner in which they produce their metadata. Partnerships and collaborations with their national statistical agency would build capacity within such organisations and promote institutional knowledge on the requirements for such activities which are not related to their core expertise. Whilst intra-BRICS collaboration would be beneficial to share knowledge across the BRICS, intra-country collaboration could be more valuable to the promotion of data quality.

In support of the data accuracy dimension, it was found that most government departments are independently audited. It was not determined if the data collection processes for each survey and their results were also audited. It is recommended that such audit reports are reviewed. It may be found that the audit reports expose greater information about the data collection processes than were included in the released metadata. Such information could influence the data user in terms of the associated level of data quality, especially in instances where the released metadata was assessed negatively.

In aid of improving data quality across the BRICS, it is recommended that the BRICS statistical agents form a BRICS statistics committee which meets regularly to formulate a strategy regarding how to go about standardising the data collection process. This is especially relevant for data collection instruments which are similarly themed. The committee’s deliberation on the promotion and adoption of a common data quality framework will guide the collection processes of the member countries and possibly signal a shift towards possible uniformity in data collection practices. The dimensions and findings of this study can be useful in determining relevant data quality dimensions that require assessment.

In heeding the call of the BRICS Heads of State it is recommended that the BRICS Think Tank Council institute research projects which examine the broad social challenges across the countries in a manner similar to this study but at a much broader scale. The findings of such a study would be useful to identify common social concerns, identify the existing data gaps which do exist and thereafter possibly institute common data collection strategies informed by an agreed set of standards which informs their data collection and data quality protection practices. Introducing such measures would go a long way to bring data collection practices in line amongst the BRICS. Furthermore, whilst the intention of the BRICS is to promote cross country collaboration, similar efforts should be adopted within the countries as well. Collaboration between data providers which produce high data quality efforts to those who do not perform strongly, can help inculcate data quality processes within such organisations and thereby foster improved data quality processes over time.

9.6. Areas for future research

Firstly, this study can be expanded to include the needs of Russia and China.

Secondly, as recommended above, the approach of the study to assess real freedoms and their supporting Capability Set could be expanded to cover the wider set of social challenges affecting the BRICS.

Thirdly, the assessment of data quality should be broadened to not only focus on the latest released dataset, but assess previous iterations of the data collection instrument as well. This would inform the data user of the reliability of older versions of the survey as well.

Bibliography

Alkire, S. et al., 2009. *An Introduction to the Human Development and Capability Approach* 1st ed. S. Deneulin & L. Shahani, eds., Earthscan. Available at: http://www.idrc.ca/EN/Resources/Publications/openebooks/470-3/index.html#page_22.

Alkire, S., 2002. The Capability Approach and Human Development. *Wadham College and and Queen Elizabeth House ….* Available at: http://www.ophi.org.uk/wp-content/uploads/SS12-CA-Intro.pdf [Accessed May 9, 2016].

Alkire, S. et al., 2008. Using the Capability Approach. *E-Bulletin of the Human Development and Capability Association Number 12, October 2008*, (12), pp.2–3. Available at: http://www.uio.no/studier/emner/sv/oekonomi/ECON4270/h09/Maitreyee12_October08.pdf.

Alkire, S., 2007. Why the Capability Approach? *Journal of Human Development*, 6(1), pp.37–41. Available at: http://www-tandfonline-com.ez.sun.ac.za/doi/abs/10.1080/146498805200034275.

Arndt, C. & Oman, C., 2007. *Uses and Abuses of Governance Indicators*, Available at: http://www.cipe.org/sites/default/files/publication-docs/043007.pdf.

Aro, T. et al., 2011. *Assessment of Learning Disabilities: Cooperation between teachers, psychologist and parents* 1st ed. T. Aro & T. Ahonen, eds., Available at: /citations?view_op=view_citation&continue=/scholar?hl=en&start=20&as_sdt=0,5&scilib=1&citilm=1&citation_for_view=-8N_dQUAAAAJ:yFnVuubrUp4C&hl=en&oi=p\n/citations?view_op=view_citation&continue=/scholar?hl=en&as_sdt=0,5&scilib=1&citilm=1&citation_for_view=.

Arruda, P.L. de et al., 2015. Educational Systems of the BRICS countries : preliminary findings of a comparative , present and future time , adequacy analysis . *7th BRICS Academic Forum*, pp.1–24. Available at: http://www.nkibrics.ru/system/asset_docs/data/5568/7b19/6272/693b/d15e/0000/original/Pedro_Arruda_Session9.pdf?1432910617.

Askham, N. et al., 2013. *THE SIX PRIMARY DIMENSIONS FOR DATA QUALITY ASSESSMENT : Defining Data Quality Dimensions*, Available at: http://www.enterprisemanagement360.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf.

Banerjee, S. et al., 2011. Under-nutrition among adolescents: A survey in five secondary schools in rural Goa. *National Medical Journal of India*, 24(1), pp.8–11.

Blum, R., 2007. *Best practices: Building Blocks for Enhancing School Environment*, Baltimore, Maryland. Available at: http://www.jhsph.edu/mci.

Bongani, M.I., 2014. *Investigating the Causes of Learner Dropout At Secondary Schools in Johannesburg South, Gauteng*. University of South Africa. Available at: http://uir.unisa.ac.za/bitstream/handle/10500/18722/dissertation_mnguni_ib.pdf?sequence=1.

Branson, N., Hofmeyr, C. & Lam, D., 2014. Development Southern Africa Progress through school and the determinants of school dropout in South Africa Progress through school and the determinants of school dropout in South Africa. *Development Southern Africa*, 31(1), pp.106–126. Available at: http://www.tandfonline.com/loi/cdsa20\nhttp://dx.doi.org/10.1080/0376835X.2013.853610\nhttp://www.tandfonline.com/.

Brazil Ministry of Education, Brazil's Ministry of Eduation.

BRICS, 2014. *BRICS Sixth Summit: Fortaleza Declaration and Action Plan*, Broome, J., 1993. John Broome Review of Inequality Rexamined. *The Royal Economic Society, Royal Economic Journal*, 103(419), pp.1067–1069. Available at: http://www.jstor.org/stable/2234727.

Brown, B.A., 2010. Social hostility and the "dropout" syndrome: Leadership assisting youths' re-entry into school? *Educational Review*, 62(1), pp.53–67. Available at: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=psyc7&NEWS=N&AN=2010-07522-003.

Burton, P. & Leoshut, L., 2013. *School Violence in South Africa Results of the 2012 National School Violence Study*, Hansaprint. Available at: www.cjcp.org.za.

Cardoso, A.R. & Verner, D., 2006. *School Drop-Out and Push-Out Factors in Brazil: The Role of Early Parenthood, Child Labor, and Poverty*, Available at: http://papers.ssrn.com/soL3/papers.cfm?abstract_id=955862.

Carson, C.S., 2000. *What Is Data Quality? A Distillation of Experience*, Available at: http://www.thecre.com/pdf/imf.pdf.

Central Advisory Board of Education, 2004. *Universalisation of Secondary Education*, Available at: http://mhrd.gov.in/sites/upload_files/mhrd/files/document-reports/universalisation.pdf.

Ceri, S., Cochrane, R.J. & Widom, J., 2000. Practical Applications of Triggers and Constraints: Successes and Lingering Issues. In *VLDB '00 Proceedings of the 26th International Conference on Very Large Data Bases*. Morgan Kaufmann Publishers Inc., pp. 254–262. Available at: http://www.vldb.org/conf/2000/P254.pdf.

Chinyoka, K., 2014. Causes of school drop out among ordinary level learners in a resttlement area in Masvingo, Zimbabwe. *Journal of Emerging Trends in Educational Research and Policy Studies*, 5(3), pp.294–300.

Chiumia, S., 2014. Is Zimbabwe's unemployment rate 4%, 60% or 95%? Why the data is unreliable.

Chugh, S., 2011. *Dropout in Secondary Education: A Study of Children Living in Slums of Delhi*, Available at: http://www.nuepa.org/Download/Publications/Occasional Paper No. 37.pdf.

Clark, D.A., 2005. The Capability Approach: Its Development, Critiques and Recent Advances. *Economics Series Working Papers*, p.18. Available at: http://ideas.repec.org/p/oxf/wpaper/gprg-wps-032.html.

Comim, F., 2001. Operationalizing Sen's Capability Approach. In *Justice and Poverty: examining Sen's Capability Approach*. pp. 1–16. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.110.4430&rep=rep1&type=pdf.

Comim, F., Qizilbash, M. & Alkire, S., 2008. *The Capability Approach: Concepts, Measures and Applications* F. Comim, M. Qizilbash, & S. Alkire, eds., Cambridge University Press. Available at: http://www.cambridge.org/sm/academic/subjects/economics/economic-development-and-growth/capability-approach-concepts-measures-and-applications.

Craveiro, G. da S., Santana, M.T. de & Albuquerque, J.P. de, 2013. Assessing Open Government Budgetary Data in Brazil. In *ICDS 2013, The Seventh International Conference on Digital Society*. pp. 20–27. Available at: http://www.thinkmind.org/index.php?view=article&articleid=icds_2013_1_40_10183.

Cummins, R., 1996. The Domains of Life Satisfaction: An Attempt to Order Chaos. *Social Indicators Research*, 38(3), pp.303–328. Available at: http://www.jstor.org.ez.sun.ac.za/stable/27522935.

Demographic and Health Survey Programme, Demographic and Health Survey Programme. Available at: http://dhsprogram.com/[Accessed August 16, 2016].

Department of Basic Education, Department of Basic Education EMIS -Education

Management Information Systems.

Department of Basic Education, 2011. *Report on Dropout and Learner Retention Strategy to Portfolio Committee on Education*, Available at: http://www.education.gov.za/Portals/0/Documents/Reports/REPORT ON DROPOUT AND ETENTION TO PORTFOLIO COMMITTEE JUNE 2011.pdf?ver=2015-03-20-120521-617.

Dieltiens, V. & Meny-Gilbert, S., 2009. School drop-out: Poverty and patterns of exclusion. *South African Child Gauge*, 2008/2009, pp.46–49. Available at: http://www.ci.org.za/depts/ci/pubs/pdf/general/gauge2008/part_two/exclusion.pdf.

District Information System for Education (DISE), District Information System for Education (DISE). Available at: http://www.dise.in/[Accessed January 20, 2015].

Dong, X.L., Berti-Equille, L. & Srivastava, D., 2009. Integrating conflicting data: the role of source dependence. In *Proceedings of the VLDB Endowment*. pp. 550–561. Available at: http://dl.acm.org/citation.cfm?id=1687627.1687690.

Dubey, M., 2010. The Right of Children to Free and Compulsory Education Act, 2009: The Story of a Missed Opportunity. *Social Change*, 40(1), pp.1–13.

Eppler, M.J., 2006. *Managing information quality: Increasing the value of information in knowledge-intensive products and processes* 2nd Editio., Springer.

Fan, W. et al., 2012. Towards certain fixes with editing rules and master data. *VLDB Journal*, 21(2), pp.213–238. Available at: http://www.vldb2010.org/proceedings/files/papers/R15.pdf.

Fleisch, B., Shindler, J. & Perry, H., 2012. Who is out of school? Evidence from the Statistics South Africa Community Survey. *International Journal of Educational Development*, 32(4), pp.529–536. Available at: http://dx.doi.org/10.1016/j.ijedudev.2010.05.002.

Flisher, A.J. et al., 2010. Substance use and psychosocial predictors of high school dropout in Cape Town, South Africa. *Journal of Research on Adolescence*, 20(1), pp.237–255.

Franks, P. & Kunde, N., 2006. Why Metadata Matters. *The Information Management Journal*, (October). Available at: http://www.arma.org/bookstore/files/Franks-Kunde1.pdf.

Ge, M. et al., 2013. *Handbook of Data Quality* S. Sadiq, ed., Springer.

Germano, E. & Takaoka, H., 2012. Uma análise das dimensões da qualidade de dados em projetos de dados governamentais abertos. In *V Congresso Consad De Gestão Pública*. p. 22. Available at: http://repositorio.fjp.mg.gov.br/consad/handle/123456789/788.

Gouda, S. & Sekher, T. V, 2014. Factors Leading to School Dropouts in India: An Analysis of National Family Health Survey-3 Data. *IOSR Journal of Research & Method in Education*, 4(6), pp.75–83. Available at: http://www.iosrjournals.org/iosr-jrme/papers/Vol-4 Issue-6/Version-3/K04637583.pdf.

Government of India Department of Science and Technology, 2012. *National Data Sharing and Accessibility Policy-2012*, Available at: http://ogpl.gov.in/NDSAP/NDSAP-30Jan2012.pdf.

Government of India Ministry of Home Affairs Office of the Registrar General & Census Commissioner, 2011. Census of India 2011.

Government of India Ministry of Home Affairs Office of the Registrar General & Census Commissioner, Census of India 2011. Available at: http://censusindia.gov.in/[Accessed August 24, 2011].

Government of India Ministry of Human Resource Development, Government of India Ministry of Human Resource Development Statistics. Available at: http://mhrd.gov.in/statist [Accessed May 3, 2016].

Government of India Ministry of Human Resource Development, 2004. *Manual for Planning and Appraisal*,

Government of India Ministry of Statistics and Programme Implementation, About the

Ministry. Available at: http://mospi.nic.in/Mospi_New/site/inner.aspx?status=2&menu_id=5 [Accessed June 12, 2016a].

Government of India Ministry of Statistics and Programme Implementation, 2005. *Chapter 37 - Crime Statistics*,

Government of India Ministry of Statistics and Programme Implementation, Government of India Ministry of Statistics and Programme Implementation. Available at: http://mospi.nic.in/Mospi_New/site/home.aspx [Accessed May 3, 2016b].

Government of India Ministry of Statistics and Programme Implementation, 2009. *Statistical System in India*, Available at: http://mospi.nic.in/Mospi_New/upload/Statistical_System_23nov09_final.pdf.

Government of India National Statistical Commission, 2011. *Report of the Committee on Data Management Government of India*, Available at: http://mospi.nic.in/Mospi_New/upload/finalreportonData management01082011.pdf.

Graeff-Martins, A.S. et al., 2006. A package of interventions to reduce school dropout in public schools in a developing country: A feasibility study. *European Child and Adolescent Psychiatry*, 15(8), pp.442–449.

Grant, M. & Hallman, K., 2008. Pregnancy-related school dropout and prior performance in KwaZulu-Natal, South Africa. *Studies in Family Planning*, 39(4), pp.369–382.

Griffin, E., 1998. Hierarchy of Needs of Abraham Maslow. In *A First Look at Communication Theory*. McGraw-Hill, pp. 235–246. Available at: http://www.afirstlook.com/docs/hierarchy.pdf.

Griffith-Jones, S., 2014. A brics development bank: a dream coming true? *UN conference on trade and development*, (215). Available at: http://unctad.org/en/PublicationsLibrary/osgdp20141_en.pdf.

Gustafsson, M., 2011. The when and how of leaving school: The policy implications of new evidence on secondary schooling in South Africa. *Stellenbosch Economic Working Papers*.

Health Information and Quality Authority, 2011. *International Review of Data Quality*, Available at: https://www.hiqa.ie/system/files/HI-International-Review-Data-Quality.pdf.

Heystek, J. & Terhoven, R., 2015. Motivation as critical factor for teacher development in contextually challenging underperforming schools in South Africa. *Professional Development in Education*, 41(4), pp.624–639. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84906531473&partnerID=40&md5=50e9f5f39679e3304c4f74539d23afe6.

Human Development & Capability Association, Human Development & Capability Association, HDCA History and Mission. Available at: https://hd-ca.org/about/hdca-history-and-mission [Accessed May 10, 2016].

Human Development Sector Management Unit, 2010. *Achieving World Class Education in Brazil : The Next Agenda*, Available at: http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=7290-achieving-world-pdf&Itemid=30192.

Human Sciences Research Council, HSRC Research Data. Available at: http://datacuration.hsrc.ac.za/.

Human Sciences Research Council, South African Social Attitudes Survey (SASAS) 2012. Available at: http://curation.hsrc.ac.za/Dataset-407.phtml [Accessed August 27, 2016b].

Human Sciences Research Council, Trends in Mathematics and Science Study. Available at: http://curation.hsrc.ac.za/Dataset-382.phtml [Accessed August 26, 2016c].

Hunt, F., 2008. *Dropping out from school: A cross country review of literature*, Consortium for Research on Educational Access, Transitions and Equity. Available at: http://www.create-rpc.org/pdf_documents/PTA16.pdf.

Husain, Z. & Chatterjee, A., 2009. Primary Completion Rates across Socio-Religious Communities in West Bengal. *Economic and Political Weekly*, 44(15), pp.59–67. Available at: http://www.jstor.org/stable/40279135.

INEP, National Literacy Assessment/Avaliação Nacional da Alfabetização (ANA) 2014. Available at: http://sitio.educacenso.inep.gov.br/web/saeb/ana [Accessed August 19, 2016].

Instituto Brasileiro de Geografia e Estatística, Instituto Brasileiro de Geografia e Estatística Statistical Portal. Available at: http://www.ibge.gov.br/english/[Accessed May 3, 2016].

Instituto Brasileiro de Geografia e Estatística - IBGE, 2014. *Censo Demográfico 2010*,

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Available at: http://portal.inep.gov.br/[Accessed May 3, 2016].

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, School Census. Available at: http://portal.inep.gov.br/basica-censo [Accessed August 20, 2016a].

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP, *Taxas de rendimento escolar*, Available at: http://download.inep.gov.br/educacao_basica/educacenso/situacao_aluno/documentos/2015/taxas_rendimento_escolar.pdf.

International Monetary Fund, 2013. *The Special Data Dissemination Standard*, Available at: https://www.imf.org/external/pubs/ft/sdds/guide/2013/sddsguide13.pdf.

International Monetary Fund Statistics Department, 2003. *Data Quality Assessment Framework and Data Quality Program*, Available at: https://www.imf.org/external/np/sta/dsbb/2003/eng/dqaf.htm.

International Monetary Fund Statistics Department, 2010. *IMF' s Data Quality Assessment Framework*, Available at: http://unstats.un.org/unsd/accsub/2010docs-CDQIO/Ses1-DQAF-IMF.pdf.

Jerven, M., 2016. *Data and Statistics at the IMF: Quality Assurances for Low-Income Countries*, Available at: http://www-ieo.imf.org/ieo/files/completedevaluations/BP6_-_Data_and_Statistics_at_the_IMF—Quality_Assurances_for_Low-Income_Countries.PDF.

Jewitt, S. & Ryley, H., 2014. It's a girl thing: Menstruation, school attendance, spatial mobility and wider gender inequalities in Kenya. *Geoforum*, 56, pp.137–147. Available at: http://dx.doi.org/10.1016/j.geoforum.2014.07.006.

Juran, J.M. & Godfrey, A.B., 1998. *Juran's Quality Control Handbook* J. M. Juran & A. B. Godfrey, eds., McGraw-Hill. Available at: http://www.worldcat.org/oclc/17546189.

Karande, S., 2008. Current challenges in managing specific learning disability in Indian children. *Journal of Postgraduate Medicine*, 54, pp.75–77. Available at: http://shibboleth.ovid.com/secure/?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=emed8&AN=2008251889

http://sfx.kcl.ac.uk/kings?genre=article&atitle=Current+challenges+in+managing+specific+learning+disability+in+Indian+children&title=Journal+of+Postgraduate+Medicine&.

Karra, M. & Lee, M., 2012. Human capital consequences of teenage childbearing in South Africa. *Poppov Research Network*, (March).

Khemangkorn, V., 1999. *A Move Toward Better Data Quality: Thailand's Progress Report*, Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.202.4745&rep=rep1&type=pdf.

Kiregyera, B., 2015. *The Emerging Data Revolution in Africa* First Edit., SUN MeDIA Stellenbosch.

Kirk, J. & Sommer, M., 2006. Menstruation and body awareness: linking girls' health with girls' education. *Tropical Institute (KIT), Special on Gender and Health*, pp.1–22. Available at: http://www.wsscc.org/sites/default/files/publications/kirk-2006-menstruation-kit_paper.pdf\nhttp://www.susana.org/_resources/documents/default/2-1200-kirk-2006-menstruation-kit-paper.pdf.

Klasen, S. et al., 1998. Measuring poverty and deprivation in south africa. *Review of Income and Wealth*, 46(1), pp.33–58. Available at: http://doi.wiley.com/10.1111/j.1475-4991.2000.tb00390.x.

Kumar, P.S., Panda, A.K. & Jena, D., 2013. MINING THE FACTORS AFFECTING THE HIGH SCHOOL. *International Journal of Advanced Computer Engineering and Communication Technology*, 2(3).

Laliberté, L., Grünewald, W. & Probst, L., 2003. Data quality: A comparison of IMF's data quality assessment framework (DQAF) and Eurostat's quality definition. In ... *Assessing and Improving Statistical Quality*. pp. 1–18. Available at: http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:DATA+QUALITY:+A+COMPARISON+OF+IMF?S+DATA+QUALITY+ASSESSMENT+FRAMEWORK+(DQAF)+AND+EUROSTAT?S+QUALITY+DEFINITION#0.

Lehohla, P., 2005. *Information for supporting decision making: the role of Stats SA*, Available at: http://www.dpsa.gov.za/dpsa2g/documents/networks/research/10_2005/SG.pdf.

Lehohla, P., 2001. *The Creation of a National Statistics System – Challenges for South Africa*, Available at: http://www.stat.go.jp/english/info/meetings/iaos/pdf/lehohla.pdf.

Loshin, D., 2001. *Enterprise Knowledge Management. The Data Quality Approach*, Morgan Kaufmann.

Lund, C. et al., 2011. Poverty and mental disorders: Breaking the cycle in low-income and middle-income countries. *The Lancet*, 378(9801), pp.1502–1514. Available at: http://dx.doi.org/10.1016/S0140-6736(11)60754-X.

Lyon, M., 2008. *Assessing data quality*, Available at: http://www.bankofengland.co.uk/statistics/Documents/ms/articles/art1mar08.pdf.

Makgato, M. & Mji, A., 2006. Factors associated with high school learners' poor performance: a spotlight on mathematics and physical science. *South African Journal of Education*, 26(2), pp.253–266. Available at: http://www.ajol.info/index.php/saje/article/viewFile/25068/20738.

Meintjes, H. et al., 2009. *Child-headed households in South Africa : A statistical brief*, Available at: http://www.childrencount.org.za/uploads/brief_child_headed_households.pdf.

Mgwangqa, V. & Lawrence, L., 2008. Why do Learners Drop out of School? Learner perceptions in the Fort Beaufort District, Eastern Cape, South Africa. *Commonwealth Youth and Development*, 6(2). Available at: http://reference.sabinet.co.za.ez.sun.ac.za/webx/access/electronic_journals/cydev/cydev_v6_n2_a3.pdf.

Minesterio do Planejamento Desenvolvimento e Gestao, 2014. Plano de Dados Abertos (PDA). Available at: http://www.planejamento.gov.br/tema/governo-aberto/plano-de-dados-abertos-pda [Accessed June 11, 2016].

Ministro da Saúde, 2008. *PNDS 2006: Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher: relatório*, Available at: http://www.saude.gov.br/pnds2006.

Mukwevho, J. & Jacobs, L., 2012. The importance of the quality of electronic records management in enhancing accountability in the South African public service: A Case Study of a National Department. *Mousaion*, 30(2), pp.33–51.

Narayan, D. et al., 2000. *Voices of the poor: Crying out for change*, Oxford University Press. Available at: http://elibrary.worldbank.org/content/book/9780195216028.

National Information Standards Organization, 2004. *Understanding Metadata*, NISO Press. Available at: http://www.niso.org/publications/press/UnderstandingMetadata.pdf.

National University of Educational Planning and Administration (NUEPA), Unified District Information System for Education. Available at: http://dise.in/[Accessed August 24, 2016].

Nussbaum, M. & Sen, A., 1993. *Quality of Life* M. Nussbaum & A. Sen, eds., Oxford Scholarship Online. Available at: http://www.oxfordscholarship.com/view/10.1093/0198287976.001.0001/acprof-9780198287971.

Nussbaum, M.C., 2000. *Women and human development : the capabilities approach*, Cambridge University Press.

Open Government Data Platform India, Open Government Data Platform India.

or Sistema Nacional de Informações de Segurança Pública (SINESP), Criminal Statistcs. Available at: https://www.sinesp.gov.br/estatisticas-publicas [Accessed August 22, 2016].

Petersen, L.J., 2010. *Parents' and educators' perceptions of factors influencing high rate of academic failure of learners in Clarke Estate Primary Schools*. University of the Western Cape. Available at: http://etd.uwc.ac.za/xmlui/bitstream/handle/11394/1950/Petersen_MED_2010.pdf?sequence=1.

Porche, M. V et al., 2011. Childhood trauma and psychiatric disorders as correlates of school dropout in a national sample of young adults. *Child Development*, 82(3), pp.982–998. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3089672/pdf/nihms193666.pdf.

Porteus, K. et al., 2000. "Out of School" Children in South Africa: An Analysis of Causes in a Group of Marginalised, Urban 7- to 15-year-olds. *Support for Learning*, 15(1), pp.8–12. Available at: http://doi.wiley.com/10.1111/1467-9604.00135.

Pritchett, L., 2004. Towards A New Consensus for Addressing the Global Challenge of the Lack of Education. *Social Science Research Network*, pp.1–72. Available at: http://ssrn.com/abstract=1112689.

Rademeyer, A., 2014. 47% quit school at Grade 10. *Beeld*.

Raju, B. & Singh, A., 2011. Educational Development in India at Elementary Level. *Indian educational review*, 49(2), pp.64–79. Available at: http://www.ncert.nic.in/publication/journals/pdf_files/indian_education_review/IER-July_2011_V49_N2.pdf.

Reddy, A.N. & Sinha, S., 2010. *School dropouts or pushouts? Overcoming barriers for the Right to Education*, Create 2010. Available at: http://www.create-rpc.org/pdf{_}documents/PTA40.pdf.

Republic of South Africa Government Gazette, 2012. *National Qualification Framework Act: DHET 002 - Data quality standard for surveys*, Republic of South Africa. Available at: http://www.gov.za/sites/www.gov.za/files/35560_gen610.pdf.

Robeyns, I., 2011. *The Capability Approach* Summer 201. E. Zalta, ed., Available at: http://plato.stanford.edu/entries/capability-approach/.

Roe, R.L., 1991. Valuing Student Speech : The Work of the Schools as Conceptual Development Valuing Student Speech: The Work of the Schools as Conceptual Development. *California Law Review*, 79(5). Available at: http://scholarship.law.berkeley.edu/californialawreview/vol79/iss5/3/.

Sabates, R. et al., 2010. *School Dropout: Patterns, Causes, Changes and Policies*, Available at: http://unesdoc.unesco.org/images/0019/001907/190771e.pdf\npapers2://publication/uuid/35234575-C690-44CE-AF2B-2A739AA78171.

Saito, M., 2003. Amartya Sen ' s Capability Approach to Education : A Critical Exploration. *Journal of Philosophy of Education*, 37(1), pp.17–33.

Sajjad, H. et al., 2012. Socio-Economic Determinants of Primary School Dropout : Evidence from South East Delhi , India. *European Journal of Social Sciences*, 30(3), pp.391–399. Available at: https://www.researchgate.net/profile/Haroon_Sajjad/publication/266262739_Socio-Economic_Determinants_of_Primary_School_Dropout_Evidence_from_South_East_Delhi_India/links/542bb8590cf29bbc126a967d.pdf.

Sen, A., 2000. A Decade of Human Development. *Journal of Human Development*, 1(1), pp.17–23.

Sen, A., 2004. Capabilities, Lists, and Public Reason: Continuing the Conversation. *Feminist Economics*, 10(3), pp.77–80.

Sen, A., 1985. *Commodities and Capabilities*, Amsterdam: North-Holland. Available at: http://www.amazon.com/Commodities-Capabilities-Amartya-Sen/dp/0195650387/ref=sr_1_1?s=books&ie=UTF8&qid=1310679705&sr=1-1.

Sen, A., 1983. Development: Which Way Now? *The Economic Journal*, 93(372), pp.745–762. Available at: http://www.jstor.org.ez.sun.ac.za/stable/pdf/2232744.pdf.

Sen, A., 2003. Development as Capability Expansion. *Readings in Human Development: Concepts, Measures and Policies for a Development Paradigm*, pp.41–58.

Sen, A., 1999. *Development as Freedom* A. Knopf, ed., Alfred A. Knopf Inc.

Sen, A., 1992. *Inequality Reexamined*, Oxford University Press.

Sen, A., 2009. *The Idea of Justice*, Harvard University Press.

Sen, A. et al., 1987. *The Standard of Living: The Tanner Lectures* 1st ed. G. Hawthorn, ed., Cambridge University Press. Available at: http://doi.wiley.com/10.1111/j.1468-0149.1988.tb02065.x.

Shah, H., 2005. Pyscho-social aspects of academic failure in Children. *Health Administrator*, 17(1), pp.34–37. Available at: http://medind.nic.in/haa/t05/i2/haat05i2p34.pdf.

Shahidul, S.M. & Karim, A.H.M.Z., 2015. Factors contributing to school dropout among the girls: a review of literature. *European Journal of Research and Reflection in Educational Sciences*, 3(2), pp.25–36.

Sidda, N.K., 2009. *A framework for the management of spatial data quality information*.

Smink, J. & Reimer, M., 2015. *Rural School Dropout Issues: Implications for dropout prevention*, Available at: http://dropoutprevention.org/wp-content/uploads/2015/05/13_Rural_School_Dropout_Issues_Report.pdf.

Soares, T.M. et al., 2015. Factors associated with dropout rates in public secondary education in Minas Gerais. *Educação e Pesquisa*, 41(3). Available at: http://www.scielo.br/scielo.php?pid=S1517-97022015000300757&script=sci_arttext&tlng=en.

Sood, N., 2010. *Malnourishment Among Children in India: Linkages with Cognitive Development and School Participation*, Available at: http://www.nuepa.org/new/Download/Publications/Create/PTA 2010/PTA25.pdf.

South African Institute of Race Relations, 2015. *Worrying dropout rates at school and university level –*,

South African Medical Research Council, SA MRC Overview. Available at: http://www.mrc.ac.za/[Accessed August 16, 2016a].

South African Medical Research Council, South African National Youth Risk Behaviour Survey. Available at: http://www.hsrc.ac.za/en/research-data/view/6874 [Accessed

August 18, 2016b].

South African Police Services, Crime Situation in South Africa. Available at: http://www.saps.gov.za/resource_centre/publications/statistics/crimestats/2015/crime_stats.php [Accessed August 16, 2016].

Southern Africa Labour and Development Research Unit, National Income Dynamics Study. Available at: http://www.nids.uct.ac.za/[Accessed July 17, 2016].

Spaull, N., 2015. Schooling in South Africa: How low quality education becomes a poverty trap. *South African Child Gauge*, (12), pp.34–41. Available at: http://nicspaull.com/research/.

Spaull, N., 2013. South Africa's Education Crisis: The quality of education in South Africa 1994-2011. *Cde*, 27(October). Available at: http://www.section27.org.za/wp-content/uploads/2013/10/Spaull-2013-CDE-report-South-Africas-Education-Crisis.pdf.

Srivastava, P.G., 2014. *Gender Concerns in Education*, Available at: http://www.ncert.nic.in/departments/nie/dse/activities/advisory_board/PDF/Genderconcerns.pdf.

Stanton, E., 2007. *The Human Development Index: A History*,

Statistics South Africa, 2008. *South African Statistical Quality Assessment Framework (SASQAF)* 1st Editio., Available at: http://www.statssa.gov.za/standardisation/SASQAF_Edition_2.pdf.

Statistics South Africa, 2010a. *South African Statistical Quality Assessment Framework (SASQAF) (1st Edition)* 1st Editio., Available at: http://mdgs.un.org/unsd/dnss/docs-nqaf/SouthAfrica-SASQAF_OpsGuidelines_Edition_1.pdf [Accessed December 3, 2015].

Statistics South Africa, 2010b. *South African Statistics Quality Assessment Framework (SASQAF)* 2nd Editio., Available at: http://www.statssa.gov.za/standardisation/SASQAF_Edition_2.pdf.

Statistics South Africa, 2010c. *Statistics South Africa Strategic Plan 2010/2011-2014/2015*, Available at: http://www.gov.za/sites/www.gov.za/files/StatsSA_strategy_plan_2010-2015_07042010.pdf.

Streeten, P., 2000. A Review Essay on Amartya Sen, Development as Freedom. *Population and Development Review*, 26(1).

Sugden, R., 1994. Review of the Quality of Life. *The Economic Journal*, 104(425), pp.950–953.

Tejay, G., Dhillon, G. & Chin, A., 2006. Data quality dimensions for information systems security: A theoretical exposition. *Security Management, Integrity, and Internal Control in ...*, (1995). Available at: /citations?view_op=view_citation&continue=/scholar?hl=it&start=10&as_sdt=0,5&scilib=1&citilm=1&citation_for_view=1DBVPZUAAAAJ:Wp0gIr-vW9MC&hl=it&oi=p.

The DHS Program, Demographic and Health Surveys. Available at: http://dhsprogram.com/what-we-do/survey/survey-display-355.cfm [Accessed August 24, 2016].

The World Bank, The World Bank Data Methodologies.

Thurlow, M.L., Sinclair, M.F. & Johnson, D.R., 2002. Students with Disabilities who drop out of school, implications for policy and practice. *National Center on Secondary Education and Transition, Issue Brief*, 1(2), pp.1–8. Available at: http://www.ncset.org/publications/issue/NCSETIssueBrief_1.2.pdf.

Topkin, B., Roman, N.V. & Mwaba, K., 2015. Attention Deficit Disorder (ADHD): Primary school teachers' knowledge of symptoms, treatment and managing classroom behaviour. *South African Journal of Education*, 35(2), pp.1–8. Available at: http://www.scielo.org.za/scielo.php?script=sci_arttext&pid=S0256-

01002015000200013&lng=en&nrm=iso&tlng=en.

Tramontina, S. et al., 2001. School dropout and conduct disorder in Brazilian elementary school students. *Canadian Journal of Psychiatry*, 46(10), pp.941–947.

Ul Haq, M., 1995. *Reflections on Human Development*, Oxford University Press.

Ul Rehman, M., 2009. Mutee-ul-Rehman Review of Idea of Justice. *The Pakistan Development Review*, 48(2), pp.173–175. Available at: http://www.jstor.org/stable/41260921.

UNDP Human Development Reports, What is human development? *UNDP Human Development Reports*. Available at: http://hdr.undp.org/en/humandev [Accessed May 12, 2016].

United Nations Development Programme, 1990. *Human Development Report 1990* B. Ross-Larson, E. Hanlon, & American Writing Corporation, eds., Oxford University Press. Available at: http://hdr.undp.org/sites/default/files/reports/219/hdr_1990_en_complete_nostats.pdf.

United Nations Statistics Division, Fundamental Principles of National Official Statistics.

Unterhalter, E., Vaughan, R. & Walker, M., 2015. The Capability Approach and Education. *Prospero*, 1(November). Available at: https://www.nottingham.ac.uk/educationresearchprojects/documents/developmentdiscourses/rpg2008walkermclean9.pdf.

Vannan, E., 2001. Quality Data — An Improbable Dream ? *Educause Quarterly*, (1), pp.56–58.

Vermuelen, R., 2014. A Capability Approach Towards the Quality of Education. *Igarss 2014*, (1).

De Waal, E., Mestry, R. & Russo, C., 2011. Religious and cultural dress at school: a comparative perspective. *PER*, 14(6). Available at: http://www.nwu.ac.za/sites/www.nwu.ac.za/files/files/p-per/issuepages/2011volume14no6/2011(14)6deWaaleaReligious&CulturalDressDOC.pdf.

Walker, M., 2005a. Amartya Sen's capability approach and education. *Educational Action Research*, 13(1), pp.103–110. Available at: http://www.tandfonline.com/doi/abs/10.1080/09650790500200279.

Walker, M., 2005b. Amartya Sen ' s capability approach and poverty analysis. *Educational Action Research*, 13(1), pp.103–110.

Wallace, L., 2004. Freedom as Progress. *People in Economics*, (September), pp.4–7.

Wand, Y. & Wang, R.Y., 1996. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), pp.86–95. Available at: http://web.mit.edu/tdqm/www/tdqmpub/WandWangCACMNov96.pdf.

Wang, R.Y.W. & Strong, D.M., 1996. Beyond Accuracy : What Data Quality Means to Data Consumers. *Management Information Systems*, 12(4), pp.5–34.

Wells, T., Sen's Capability Approach. *Internet Encyclopedia of Philosophy*. Available at: http://www.iep.utm.edu/sen-cap/.

World Food Program, 2010. *Feed Minds, Change Lives: School Feeding , the Millennium Development Goals and Girls' Empowerment*, Available at: http://www.un.org/en/ecosoc/innovfair2011/docs/wfp.pdf.

Master's Thesis

Identifying Social Indicators for the BRICS using public data

An investigation of the School Dropout Phenomenon in Brazil, India and South Africa

Appendices 1 - 5

Krish Chetty

Student Number: 17457181

Master of Philosophy (Information and Knowledge Management)

March 2017

Supervisor

Heidi van Niekerk

Appendix 1

Applicable Data Sources Per Functioning

Functioning	Theme	Revised Sub-Theme	Brazil	India	South Africa
Being physically well	Malnutrition and Hunger	Access to feeding programmes	PNDS 2006	DISE	🟥
		Number of learners affected by malnutrition	PNDS 2006	🟥	YRBS, TIMSS 2011, DHS 2003
		Nutrition intake of learners	PNDS 2006	DHS	DHS 2003
		The effect of extreme poverty and hunger	PNDS 2006	🟥	SASAS 2012, CS2016
	Menstruation and Female Maturation	Gender biased communities	PNDS 2006	Census 2011, DHS 2008-2015	SASAS 2012, DHS2003
		Access to female sanitary products	🟥	🟥	🟥
		Understanding female maturation	PNDS 2006	DHS 2008-2015	🟥
		Cost of female sanitary products	🟥	🟥	🟥
		Access to adequate sanitation in schools	School Census	🟥	🟥
		Provision of medication to counter menstruation discomfort	🟥	🟥	🟥
	Teenage Pregnancy	Provision of sex education	School Census	🟥	🟥
		Pregnant below 20 years of age	PNDS 2006	DHS 2008-2015	GHS 2015, DBE ASS, NIDS, Census 2011, CS2016
		Accessibility and knowledge of contraception	PNDS 2006	DHS 2008-2015	YRBS, DHS2003
		Late School Entry	🟥	DISE	DBE ASS

Functioning	Theme	Revised Sub-Theme	Brazil	India	South Africa
Being financially secure	Financial Security	Low Household Income	🟥	Census 2011, DHS	SASAS 2012, NIDS, Census 2011, IES2010
		Cost of school fees and other resources	🟥	🟥	DBE ASS, Schools Master-List, NIDS, Census 2011, IES2010, GHS2014
		Available space for home study and necessary household resources	IBGE Census	🟥	🟥
		Socio-economic status of the learner, school and community	Employment Status, Household Income	🟥	Schools Master List, Census 2011
		Cost of school fees compared to household income	🟥	🟥	🟥
	State funded programmes	Provision of state grants targeted at poor households	School Census, IBGE Census	🟥	DBE ASS, SASAS 2012, NIDS, GHS 2015, DHS2003
		Provision of free schooling	🟥	🟥	Schools Master List, NIDS
	Opportunity cost of school attendance	Opportunity cost of school attendance in the form of employment income	🟥	Census 2011	🟥
	Access to basic services	Access to basic services	PNDS 2006, IBGE Census	DHS	SASAS 2012, Census 2011, IES 2010
Being Mentally Well	Psychological factors	Social exclusion of poorer learners	🟥	🟥	🟥
		Shame of poverty	🟥	🟥	🟥
		Impatient teachers	🟥	🟥	🟥
		Disinterested learners	🟥	🟥	🟥

Functioning	Theme	Revised Sub-Theme	Brazil	India	South Africa
		Difficulty in learning complex subject matter	🟥	🟥	🟥
		Teachers skills not sufficient for complex subject matter	School Census	🟥	🟥
	Neurological factors	Social exclusion of poorer learners	🟥	🟥	🟥
		Shame of poverty	🟥	🟥	🟥
		Impatient teachers	🟥	🟥	🟥
		Disinterested learners	🟥	🟥	🟥
		Difficulty in learning complex subject matter – teaching to memorise	🟥	🟥	🟥
		Teachers skills not sufficient for complex subject matter	🟥	🟥	🟥
		Support for attention disorder symptoms	🟥	🟥	🟥
		Support for learning disability symptoms	School Census	🟥	🟥
	Emotional factors	Access to psychological services	🟥	🟥	🟥
		Depression	PNDS 2006	🟥	🟥
		Stress	PNDS 2006	🟥	SASAS 2012
		Are there learning deficiencies experienced in reading, spelling, writing, comprehension, mathematics, problem-solving and attention	🟥	🟥	🟥
		What are the experiences of trauma affecting the learner, i.e. crime, violence, abuse (domestic, family, sexual)	🟥	🟥	SAPS
Traveling to school safely and conveniently	Distance to school	Distance, time learner commutes to school	School Census	DISE	NIDS, Census 2011
		Available Access Road to school	🟥	DISE	🟥
		Urban/Rural location of the school	🟥	🟥	🟥
		Available transportation services	School Transportation	🟥	NIDS

Functioning	Theme	Revised Sub-Theme	Brazil	India	South Africa
	Provision of transportation	Cost of transportation	🟥	🟥	NIDS, IES 2010
	Sexual harassment of female learners	Female learners suffer sexual harassment in commute to school	🟥	🟥	🟥
Being in a conducive home learning environment	Negative attitude to learning in the household, community	Education attainment of the head of household and associated community	IBGE Census	Census 2011, DHS	NIDS, Census 2011, GHS 2015, DHS2003
		Negative Environment	PNDS 2006	🟥	🟥
		Head of household's opinion of education value	🟥	🟥	🟥
		Personal and Community's opinion of education value	🟥	🟥	SASAS 2012
	Abuse in the household and community	Physical and verbal abuse	🟥	DHS 2008-2015	🟥
		Substance (drug and alcohol) abuse within the household and community	🟥	DHS 2008-2015	🟥
		Conduct disorders of learners	🟥	🟥	🟥
	Child Headed Households	Child headed households	IBGE Census	Census 2011	NIDS, Census 2011
		Migrant caregivers	🟥	🟥	NIDS
		Household income of child headed households	IBGE Census	🟥	🟥
Being Taught in Suitable Infrastructure with necessary resources	Improve Physical Infrastructure	Ensure that the schools are of a suitable size	🟥	DISE	DBE NEIMS, TIMSS 2011, NIDS
		School infrastructure maintenance	School Census	🟥	DBE NEIMS, TIMSS 2011
		State of disrepair of the school	School Census	DISE	DBE NEIMS
	Effects of Physical Infrastructure	School's access to basic services	School Census	DISE	DBE NEIMS
		Classrooms mixed by grade and subject area due to limited space	🟥	🟥	DBE ASS

Functioning	Theme	Revised Sub-Theme	Brazil	India	South Africa
		Reputation of the school to provide opportunities	🟥	🟥	🟥
	Facilities and Resources	Available facilities within a school	School Census	DISE	DBE NEIMS
		Security of the school	🟥	DISE	DBE NEIMS
		Available learning support materials	School Census	🟥	🟥
		Available teaching aides	School Census	🟥	TIMSS 2011
		Secure community environment	🟥	🟥	YRBS
Free Expression of Opinions	Negative teacher attitude	Restrictive classroom environment	🟥	🟥	🟥
		Teachers with preconceived attitudes to learners	🟥	🟥	🟥
	Conducive school learning environment	Undisciplined learners	🟥	🟥	SASAS 2012, TIMMS 2011
		Appropriate number of teachers	🟥	DISE	DBE NEIMS, School Master-list, DBE ASS
	Restricted cultural expression	Restrictions on cultural attire	🟥	🟥	🟥
		Promotion of Cultural diversity	School Census	🟥	🟥
		Restrictions on free speech	🟥	🟥	🟥
Meaningful participation in school	Comparative academic strength of the learner	Understanding elementary level concepts	ANA 2014	Census 2011	Census 2011, GHS2014, GHS2003
		Comprehension of classwork	ANA 2014	DISE, MHRD	DBE ANA, TIMSS 2011, NIDS
	Teacher Quality	Teacher professionalism	School Census	🟥	🟥
		Pedagogical skills, training and knowledge for teaching	School Census	DISE	TIMSS 2011
		The ability to impart and instil knowledge, skills and values to the learners	🟥	🟥	TIMSS 2011
		Lack of teacher motivation	🟥	🟥	TIMSS 2011

Functioning	Theme	Revised Sub-Theme	Brazil	India	South Africa
	Teacher Motivation	Positive and negative teacher incentives	🟥	🟥	🟥
	Household and Community Environment	Learner's motivation for school work	🟥	🟥	NIDS
		Crime, violence, substance abuse and other negative environmental factors	Crime Statistics Portal	MOSPI	SAPS, YRBS, CS2016
		Socio-economic conditions	IGBE Census	DHS 2008-2015	🟥
	Discrimination against some learners	Discrimination in the household, school and community	🟥	🟥	SASAS 2012
		Report school attendance trends by Gender, Race, Caste, Creed	School Census	🟥	🟥
		Include sensitivity training pertaining to marginalised groups in the curriculum	🟥	🟥	🟥

Appendix 2

Application of the PDQAF for each survey

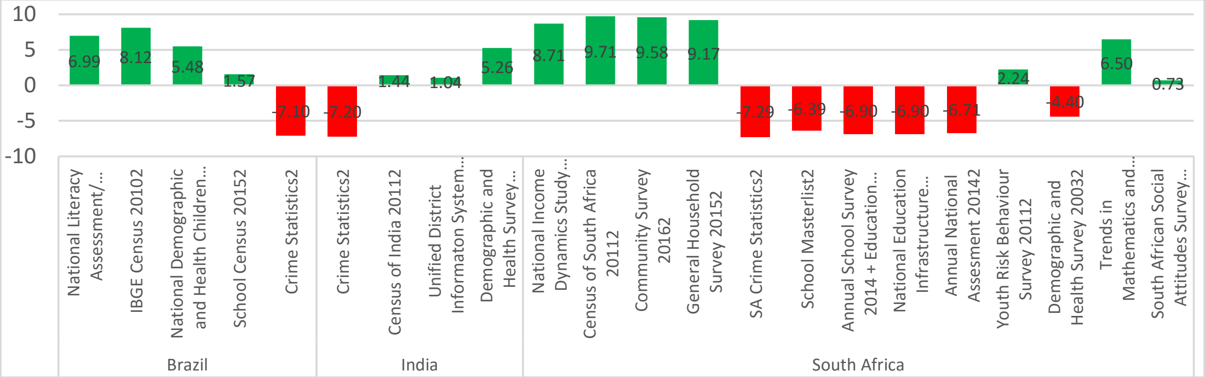


Figure 12: Data quality assessment summary of data collection instruments conducted in Brazil, India and South Africa

a. Brazilian Data Collection Instruments

Within Brazil, five data collection instruments were identified that provide data pertaining the valued set of functionings of learners whom are faced with dropping out of school. These collection instruments are discussed in the below table.

Data Provider	Collection Instrument	Score
Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)	National Literacy Assessment/Avaliação Nacional da Alfabetização (ANA) 2014	6.99
	School Census 2015	1.57
Ministry of Health	National Demographic and Health Children and Women/Pesquisa Nacional de Demografia e Saúde da Criança e da Mulher (PNDS-2006)	5.48
Brazilian Institute of Geography and Statistics (IBGE)	Census of Brazil 2010	8.12
National System of Public Security Information, Prisons and Drugs/Sistema Nacional de Informações de Segurança Pública (SINESP)	Brazil Crime Statistics 2014	-7.10

Table 57: Data Collection Instruments in Brazil

National Literacy Assessment 2014

The National Literacy Assessment which is conducted annually by Brazil's Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) is an example of a well performing organisation which pays significant attention to data quality, despite the fact that its primary goal is not data collection. The National Literacy Assessment in 2014, was found to provide data which primarily supported the learner's meaningful participation in school specifically in relation to their academic strength, the quality of teachers in schools and related to household and community factors. The specific questions and related sub-themes which were found relevant are described in Table 58 below. The National Literacy Assessment is used to produce indicators that describe the levels of literacy and numeracy in Brazil's schools as well as factors impacting the school. The 2014 data is the only set of data currently released, but INEP has packaged the data release with a wide range of supporting documentation and metadata which was vital for the dataset's stronger data quality score in comparison to other data collections (INEP n.d.).

Functioning	Theme	Sub Theme	Specific Question in National Literacy Assessment 2014
Meaningful participation in school	Comparative academic strength of the learner	Understanding elementary level concepts	Response to Student Reading Test in Portuguese
			Student response to Math Test
			Average Reading in Portuguese
			Average Writing in Portuguese
		Comprehension of classwork	Positioning student proficiency scale in Reading in Portuguese
			Positioning student proficiency scale in Mathematics
			Percentage of students with less than or equal proficiency to 425, 525, 625 points in Portuguese Reading
			Percentage of students with lower proficiency than 350, 450, 500, 600 points in Portuguese Writing
			Percentage of students with less than or equal proficiency to 425, 525, 575 points in Mathematics
	Teacher Quality	Pedagogical skills and knowledge for teaching	Teacher Training Adequacy Indicator (information relating to Group 1, for the Years EF Initials)

Functioning	**Theme**	**Sub Theme**	**Specific Question in National Literacy Assessment 2014**
	Household and Community Environment	Socio-economic conditions	Level school Socio Economic Data

Table 58: Data Available in the National Literacy Assessment 2014, relevant to School Dropouts

The National Literacy Assessment of 2014 has performed comparatively strongly to the other datasets from Brazilian due to INEP's packaging and release of the dataset. Effort was made to bundle the dataset together with a data dictionary, codebook, data that was released in multiple formats, technical documentation and key results related to the many questions included in the survey. Using the scoring system discussed earlier, the data collection instrument attained a score of 6.99 out of the rebased scale of 10. Whilst the survey achieved a perfect score in some dimensions, there were areas of concern that were identified, especially in connection with data traceability, applicability and accuracy.

In terms of the organisational dimension of data quality, specifically in regards to the manner useful metadata is provided to the data user, the dataset attained a score of 0.65. In terms of how the metadata described the contents of the dataset, it was found that the concepts were well described, the information was timely and related specifically to the 2014 release, the codebooks presented information on all the tables and fields included in the dataset and the data provided was extremely granular providing school level information and therefore also discussing each level of the Brazilian Education system geography hierarchy. When analysing whether the metadata assisted the data user to better understand the context of the dataset, various shortcomings were exposed. In describing the data quality practices of INEP, the metadata does not express how INEP measures data quality. Therefore, the data user is unaware of the data quality practices that are followed or if any do occur. However, in support of this particular indicator INEP does describe the test matrices that were performed to guide the sampling frames and also discusses the use of response theory and the correction tests that were applied to guide INEP in dealing with data quality breaches. Therefore, whilst it is clear that instituting procedures to address data faults is a concern, the metadata does not describe INEP's position on data quality and how it attempts to measure quality.

In regards to how comprehensive the metadata is, it was found that the statistical techniques required to manage the sampling framework and test responsiveness were alluded to but were not discussed in detail. Therefore, the user is unable to test the principles that were discussed without having access to the particular statistics. In terms of the norms and standards used to define the data collection procedure, no particular standards were referred to in the documentation, and therefore the data user is unable to assess whether the process follows a well-tested process of data collection. However, the dataset does perform well in terms of other focal issues related to metadata comprehensiveness. Anonymity of the target audience is well protected considering the dataset provides granular school level information but details of the learners is excluded from the dataset. In addition, the questionnaire structure is well documented within the metadata and greater detail is captured within the SPSS dataset itself, assisting the user to makes sense of the wide range of questions provided. In terms of the other indicators describing the context of the metadata, it was found that the supporting findings report was provided in detail and the layout, language and referencing was clear and understandable and all the information provided was directly connected to the current dataset under analysis. Lastly, in relation to the metadata, the structure of data was found to be discussed adequately well providing information on the physical layout of the data files as well as the specific software and hardware needs required by the data user when working with the dataset.

The next dimension of the framework describes how comprehensive the data is in meeting the user's needs. In this regard, the National Literacy Assessment scored perfectly across all indicators and focal issues describing this dimension as the data provided, matched all descriptions included within the metadata. This related to the detail of each statistical unit (survey question) included within the dataset, the structural layout was found to match the codebook layout, and all options referred to in each survey question were made available within the dataset and furthermore all the data rules were met.

In terms of data accuracy, due to the non-provision of statistical information within the metadata, the data user is unable to assess how accurate the data provided is. Therefore, the sampling and imputation techniques applied cannot be corroborated. However, the non-sampling related statistics were provided referring to the reference test matrices which were applied to test the data produced after the survey was completed. This information was published which supported the datasets stronger results in terms of the suitability of the data for reporting. This highlights a weakness in the INEP process whereby the pre-survey set up requirements are either not performed or not documented whilst post survey results are provided.

In reference to the clarity, applicability, conciseness, consistency and currency of the dataset, it was found that the terms used were defined and consistently used which supported how well the dataset is understood, in terms of applicability, the issue which was faced in analysing this dataset (and various others as well) is the assessment of the user's needs and how these needs factored into the formation of the questionnaire structure. In addition, as the user's needs are not discussed within the published metadata, one cannot assertively state whether the user's needs are considered, which again highlights how essential detailed metadata is to the data user. Whilst the content of the data is beneficial to user's understanding of the state of literacy in Brazil, a needs assessment could have improved the data provided or assisted in the presentation of the results or the manner in which the data is made available. In terms of the data being concise and consistent, both factors were found to be adequately addressed. With regards to the currency and timeliness of the data delivery to the public, it was found that the data was released 2 years after the legislated ordinance, which is consistent with international surveys on literacy. Furthermore, the preliminary results of the 2015 survey have been released which indicates the process flow of INEP is working at a consistent pace when releasing data to the public, although the organisation does not release a formal schedule of when the survey results will be released.

In terms the accessibility of the dataset, various focal issues where found to be not applicable, thus not factoring into the summation for the dimension. Firstly, data is freely made available on the INEP website. This data is extremely granular with no prohibitions. The only limitation on access to such data is the quality of the user's internet connection to download the sizable data files. The reasons for non-applicable focal points was because the data is already available and therefore users need not make requests for the available data. What was found was that the data was accessible, anonymised, the metadata described how one could access and use the data, the software requirements were explicitly discussed and the data was provided in multiple formats. The only negative point regarding this dimension is the scheduling of the data release which is not discussed by INEP on any of the communication platforms.

Lastly in regards to the integrity and traceability of the data provided, it was found that the metadata adequately expressed the manner in which data constraints were applied during data collection activities. Non-applicable data values were excluded by trained field-workers. In addition, the aggregations to the data that was applied were found to be adequate, following the reference tests that were documented in the metadata. What was noted was that the reported data in the statistical release confirmed the values when independently tested in aggregated form. A negative concern regarding the integrity of the data pertains to documenting the process that is followed when a breach in data quality is found. The metadata does not state how the data fault is address, whilst INEP briefly alludes to their internal process on the website. Such processes should be included in the reported metadata. In terms of data traceability, tracking the source of the information is not relevant as the data source in this instance is not an external system, but individual learners who were survey respondents.

In review of the INEP National Literacy Assessment, a key issue that emerges regards how crucial well-updated metadata is to the data user. The metadata needs to discuss the various data quality protections that are in place to assure the user that such processes were followed. The INEP metadata together with unorganised information on the INEP website alludes to certain practices which should rather be discussed in detail. If such information was available, the survey would have received a higher data quality score. The scores for the National Literacy Assessment for each data quality dimension are provided in the following graph (see Figure 13).

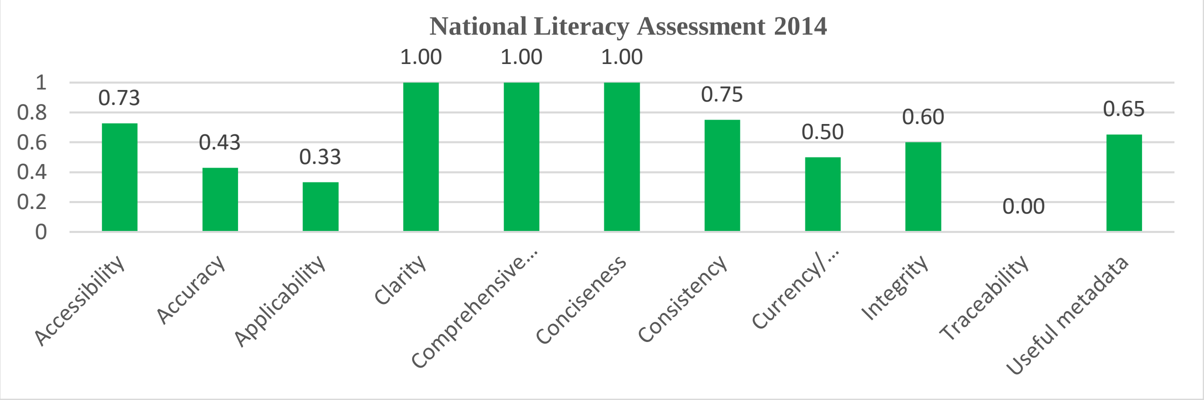


Figure 13: Data Quality Assessment of Brazil's National Literacy Assessment 2014

Brazil School Census 2015

The School Census of Brazil in 2015 as coordinated by Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) provides many questions relevant to the various functionings valued by learners in terms of the choices they make in regards to dropping out of school. The survey is conducted annually and pools the efforts of the public and private schools as well as state and municipal departments of education. The survey is very broad and targets all learners, school professionals from all schools in the country. Through the coordination of this census, Brazil is able to conduct a single questionnaire which cover many aspects of the education service as found relevant by the departments (Instituto Nacional de

Estudos e Pesquisas Educacionais Anísio Teixeira - INEP n.d.). However, although the contents of the survey are very useful and crucial to the management of the education function in the country, the score attained in this data quality assessment highlights strong concerns related to specific data quality dimensions such as integrity, accuracy and applicability. The overall score of 1.57 indicates that on the whole, the dimensions are weighed more positively than negatively but when one considers the number of negatively assessed dimensions, various improvements can be applied to the school census. The relevant content within the census which inform the functionings of the study are discussed below in Table 59.

Functioning	Theme	Sub Theme	Specific Question in School Census 2015
Meaningful participation in school	Teacher Quality	Pedagogical skills and knowledge for teaching	Has Degree in Course in Training Specific to Youth and Education
			Educational Service Specialist (AEE)
			School with pedagogical proposal by alternating training
			Didactic and pedagogical mediation offered by the school
		Teacher professionalism	Professional way of teaching corresponding to the classes of initial and continuing education courses or vocational qualification (courses FIC) articulated to EJA or concomitant; or technical courses level
			Professional Education Course - Technical
Being Mentally Well	Neurological factors	Support for learning disability symptoms	Students with disabilities, pervasive developmental disorder or high in Special Needs, Blindness, Low Vision, Deafness, Difficulty Hearing, Mental Disorder, Multiple disabilities
Free Expression of Opinions	Restricted cultural expression	Promotion of Cultural diversity	Training specific to Indigenous Education
			Training specific to ethnic relations and Afro-Brazilian history and culture and African
			Specific learning materials to meet the sociocultural diversity
			Indigenous education taught
Being physically well	Teenage Pregnancy	Provision of sex education	Training specific to Gender and Sexual Diversity
	Menstruation and Female Maturation	Understanding female maturation	Training specific to Child and Teen Rights
		Access to adequate sanitation in schools	Type of sanitation
Being financially secure	Access to basic services	Access to basic services	Water consumed by students in the school goes through a filtering process
			Source of Water Supply
			Source of Electricity Supply
			Type of sanitation
			Type of Waste Disposal
Being Taught in	Facilities and Resources	Available facilities in the school	Existing facilities in school by type of facility

Functioning	Theme	Sub Theme	Specific Question in School Census 2015
Suitable Infrastructure with necessary resources		Available learning support materials	Existing equipment in school by type of equipment
	Improve Physical Infrastructure	Ensure that the schools are of a suitable size	Number of classrooms in school
Free Expression of Opinions	Conducive school learning environment	Appropriate number of teachers	Total school staff
Traveling to school safely and conveniently	Distance to school	Distance learner commutes to school	Distance Education - Distance Education
	Provision of transportation	Available transportation services	Type of transportation provided to learners

Table 59: Data Available in the Brazil School Census 2015, relevant to School Dropouts

The positively assessed dimensions that were identified include clarity of the dataset, how comprehensive the data is and how concisely the content of the information is communicated to the data user. In terms of clarity, it was found that the survey was consistent in the manner it used terms and definitions in the broad survey. Ensuring that the data, metadata and published results were consistent, helps to promote the common interpretation and understanding of the data amongst data users. In terms of data comprehensiveness, the survey was found to adequately represent all facets described within the metadata in the actual published data. This included all statistical components that were discussed, the physical structure of the data, the manner in which the dataset presents items discussed as part of the scope, definitions and classifications of content as well as the manner in which the dataset represents the rules discussed within the metadata. In terms of conciseness, the dataset was found to be to the point in terms of each of the fields that were provided and in terms of the depth of granularity of the dataset in question.

Another positive dimension that was identified is the accessibility offered by INEP to data users, however, various focal issues highlight concerns that need to be addressed such as the data release schedule which is not published and the lack of information pertaining to the software requirements for using the dataset and in considering that the data is only made available in Excel, which is limiting to users whom prefer data access in multiple data formats. Although the detailed information is made available, the process to request data or support from INEP is not clear as no steps for requesting such support are detailed within the metadata. Furthermore, the metadata also does not reference any particular data dissemination strategy that INEP follows to ensure data access is adequately managed. On the positive side, the anonymization techniques were found to be suitable and the metadata provided adequate detail of the data structures available and how to access the data.

Three dimensions were found to be neither positive or negative in the manner data is provided to the data user. Specifically, these dimensions are consistency, currency and traceability. In terms of consistency, whilst the data values were consistently reported across the dataset when one compared the data to the metadata, it was also found that one could not evaluate how the data rules were applied to the dataset as the metadata did not discuss the data collection methodology that was followed. In terms of currency two issues emerged, whilst INEP should be commended at regular annual publications of a survey of such magnitude, the lack of any data release schedule limits data user's ability to work with the data in a timely fashion. Lastly in terms of tracing the source of the data, this dimension was found to be non-applicable as no source system was relevant to be tracked.

The negative dimensions which have the most cause for concern are integrity, accuracy, applicability and useful metadata. Data integrity was found to be a problem as the data quality constraints that may or may not have been included in the dataset are not discussed within the metadata. Thus the user cannot determine how data quality was managed during data capture or consolidation. In terms of accuracy concerns, various shortcomings in the published metadata were also identified which related to the lack of documented data sampling techniques or in terms of statistical calculations pertaining to standard errors or the imputation techniques that were employed. Furthermore, no comparative study was identified within the metadata to help verify the results of the survey. Within other data collections, it was noted that some studies employed a pre-sampling or post enumeration survey to help test the results of the data collection. No such study was referenced here or discussed in the metadata. In terms of applicability, the concern that emerged was also related to the detail provided within the metadata, which failed to discuss whether the users' requirements when using such a dataset were determined. Furthermore, neither were the public's perceptions of the dataset discussed within the metadata. Lastly, whilst many of these factors referred to the metadata, when analysing the metadata it is apparent that whilst an attempt at providing some information to the user was made, the documentation excluded information on the standards employed when defining the questionnaire, INEP's definitions of data quality and how the organisation attempted to manage data quality during the data collection phase or in the fact that the user is unable to determine the comprehensiveness of the metadata as the scope of the questionnaire is not discussed within the metadata.

In assessing all the data quality dimensions, the degree of content within the metadata drives the data quality score that is achieved. Whilst INEP makes an effort to discuss the structure and definitions of the very broad questionnaire that is rolled out annually, the organisation does not document the data quality practices that may be followed. In the absence of such details, the user must be cautious and assume that such data quality practices were not performed. If greater detail related to the data quality factors are documented within the metadata, a higher score may be achieved. The specific scores per dimension are discussed below in Figure 14.

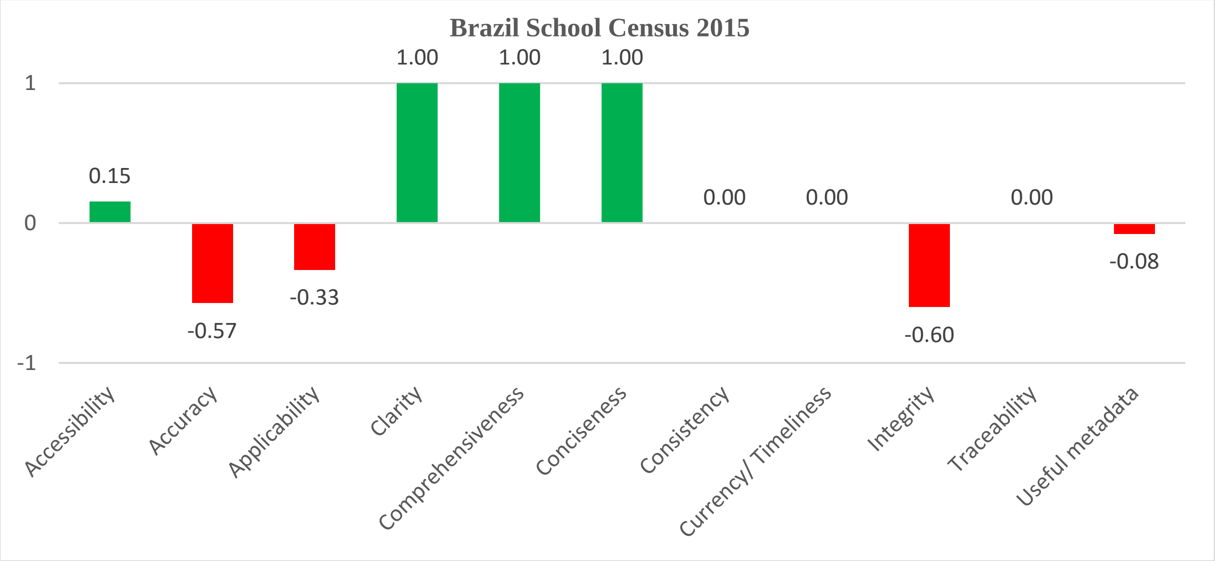


Figure 14: Data Quality Assessment of Brazil School Census 2015

National Demographic and Health Children and Women Survey 2006

The National Demographic and Health, Children and Women Survey conducted in 2006 (also referred to by the Portuguese acronym PNDS 2006) and is managed by the Brazilian Ministry of Health. The study is supported by USAID and the National Ministry is supported by various international institutions to measure a wide range of health related indicators such as population analysis and the nutrition of women and children. The study was conducted in 1986, 1996 and 2006 with each iteration broadening the scope of the survey with the 2006 survey including modules on technical and scientific advances, contraception, marriage and sexual activity and teenage pregnancy (Ministro da Saúde 2008). Each of the subject areas are found to match the broad functionings related to school learners with specific topics of the survey highlighted in Table 60. From a data quality perspective, the dataset attained a score of 5.48 performing strongly in terms of clarity, comprehensiveness, conciseness and consistency. As the survey is also supported by the International Demographic and Health Survey Programme, the questionnaire structure and data quality practices followed by the Health

Ministry have adopted standards which are internationally recognised and are key positive focal points within the public data quality assessment framework. The table below details the relevant questions per functioning, theme and sub-theme within the PNDS 2006.

Functioning	Theme	Sub Theme	Specific Question in National Demographic and Health Children and Women Survey 2006
Being physically well	Malnutrition and Hunger	Access to feeding programmes	They receive basic food basket inclusive of dairy, vegetables, and other food stuffs
		Number of learners affected by malnutrition	Index nutritional height and weight for age appropriate
		Nutrition intake of learners	Consumption of particular food stuffs over last 7 days
			M481-Indicated dose of vitamin A, in the last 06 months
		The effect of extreme poverty and hunger	Adult skipped meals
			Experiences of hunger
			Food consumption aged below 18 years
			Hunger aged below 18 years
	Menstruation and Female Maturation	Gender biased communities	Married before age of 20
		Understanding female maturation	Knowledge of times when at risk of falling pregnant
	Teenage Pregnancy	Pregnant <20 years	Reasons for falling pregnant, if below 20
		Accessibility of contraception	Knowledge of various forms of contraception
Being financially secure	Access to basic services	Access to basic services	Type of water supply
			Access to electricity at home
Being Mentally Well	Emotional Factors	Depression	Learners under 20 years that feel they have a reason for living
		Stress	Learners under 20 years that feel life has become more difficult
	Negative attitude to learning in the household, community	Negative Environment	Learners under 20 years that feel rejected by family

Table 60: Questions from the National Demographic and Health, Children and Women Survey related to functionings valued by learners

As mentioned above, the PNDS of 2006 performed strongly in terms of data clarity, comprehensiveness, conciseness, consistency and accuracy. The data quality of the survey was found to be weakest in the areas of data accessibility, applicability and currency. As found in the analysis of other Brazilian data collections, the detail provided in the published metadata is critical for achieving a higher data quality assessment score. For this particular survey, the metadata dimension achieved a score of 0.5 due to the metadata not reporting on all the structural details of the published data, not providing the geographic detail provided within the survey, the lack of documentation on the measuring of data quality and the metadata's exclusion of hardware or software requirements for accessing the data. These factors limit the metadata's strength in terms of describing the contents of the dataset, the context of the dataset and most severely against the metadata's explanation of the structure of the dataset. The metadata was found to be strong in terms of other focal issues such as definitions used within the data, the linkage of the definitions to international standards, the manner in which data quality is processed by the organisation, the data quality controls that were instituted, the comprehensiveness of the data collection methodology and the inclusion of a detailed supporting findings reports.

The weakest data quality dimension of the PNDS 2006 is the accessibility of the data, and the weak performance was due primarily to limitations of the metadata which did not provide sufficient detail about the geographic details and data file structures within the dataset which impedes the user's understanding of how to access more granular data, as the granularity is not explicitly discussed. The lack of data file descriptions, severely limited the user's understanding of whether the published data is actually useful and meaningful as the user must independently assess what each variable within the dataset actually refers to. Furthermore, the lack of a data release schedule and data dissemination strategy does not display an appreciation of the user's needs in accessing the data once it is made available. Lastly, as the data is only provided using the SPSS format, the needs of users without access to such software are excluded. Other data formats could be utilised or the Ministry could also investigate the roll out of online database tools to improve the usability of the data.

Data currency, applicability and traceability were found to be neither positive nor negative. The data currency was neutral as it was not possible to assess how long the publication of the data took due to a lack of information regarding the delays in publication and due to the infrequency of the publication. The applicability of the dataset was neutrally assessed as the metadata excluded reporting on the user's perceptions of the data, whilst the data that was included in the released dataset was found to provide only relevant information by excluding technical and unimportant data fields. The data traceability factors were found to be irrelevant as they did not publish data that was not dependent on any external source systems but rather only the individual survey responses.

Data integrity and accuracy were found to be stronger in terms of data quality with scores of 0.2 and 0.86 respectively. The primary reason for data integrity not scoring highly was due to the lack of documentation regarding how data quality breaches were addressed within the metadata whilst the documentation did adequately discuss how data constraints were applied within the electronic questionnaire. In terms of accuracy of the data, the infrequency of the data collection counted against the dataset, although other focal issues related to the discussion of data coverage methods and the suitability of the data for reporting purposes were found to be adequate.

The strength of the PNDS is highlighted in the manner the dataset meets the needs for clarity, comprehensiveness, conciseness and consistency in the manner the questionnaire is structured. Each of these factors scored highly. Clarity is assessed in terms of how well the released data corresponds with the terms and definitions used in the metadata. This process was found to be adequately represented. In terms of comprehensiveness, the scope, definitions, classifications, valuations and timelines discussed within the data corresponded strongly with the data that is released and reported within the statistical reports. The conciseness of the dataset is displayed in the detail of every field that is provided and the depth of granularity offered for each variable included in the survey. The consistency of the dataset is highlighted in how it represents data values and data rules within the dataset. All reports and data files uniformly follow a common methodology which is also consistent with the approach discussed by the international Demographic and Health Survey Programme.

Whilst the PNDS provides a wide range of crucial data variables and follows internationally recognised standards, the shortcomings of the survey relate to the limitations found in the metadata regarding the geographic structure and the lack of a detailed codebook to provide the data user the necessary information for accessing and interacting with the released datasets. These oversights together with the exclusion within the documentation of how data breaches were addressed, reduces the data quality score that was achieved. The summary of these factors is made available presented in Figure 15 below.

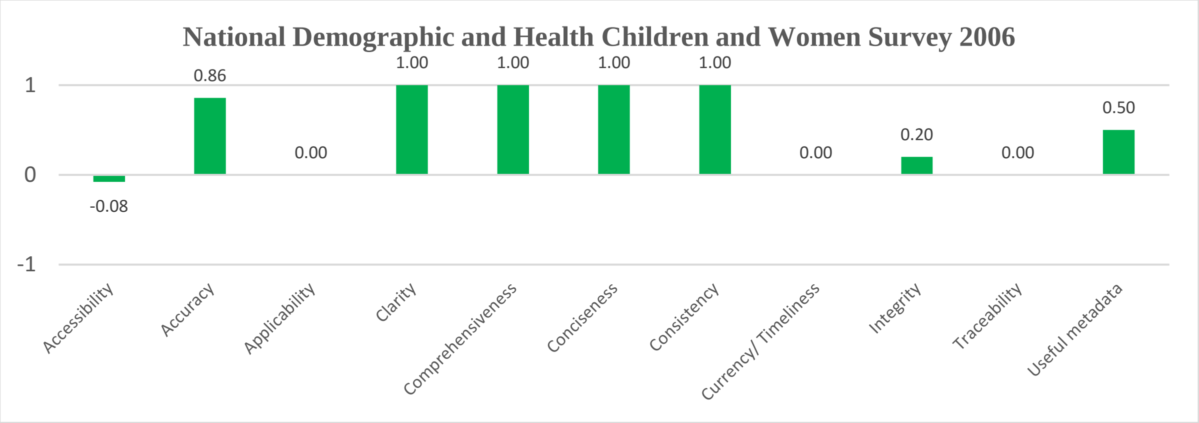


Figure 15: Data Quality Assessment of Brazil National Demographic and Health Children and Women Survey of 2006

Census of Brazil 2010

The Brazilian Census of 2010 was conducted by the Brazilian Institute of Geography and Statistics (referred to by the Portuguese acronym of IBGE) which is the official national statistics provider of the country. The census is carried out every 10 years and great care is taken to ensure reliable and accurate data is produced from which the country can base various future policy decisions. The Census targeted all households in the country and collected information regarding the characteristics of the household and the data was also used to produce social indicators regarding trends affecting municipalities in the country (Instituto Brasileiro de Geografia e Estatística - IBGE 2014). In relation to school dropouts, various questions were found to be relevant in terms of financial security and living in an environment conducive to learning. The specific questions in Table 61 below discuss which themes and sub-themes relate to the above mentioned functionings. In terms of data quality, the Brazilian Census performed the strongest in comparison to all data collection instruments from Brazil with a score of 8.12.

Functioning	Theme	Sub Theme	Specific Question in Census of Brazil 2010
Being financially secure	Financial Security	Socio-economic status of the learner, school and community	Type of Household
			Predominant Material of the External Walls of the Household
		Low Household Income	Worked earning cash, products, goods or Benefits, Benefits: Housing Food, Training
			HAD ANY PAID WORK which was TEMPORARILY AWAY?
			WORKED on the PLANTATION, breeding or FISHING, ONLY to FEED the INHABITANTS of the HOME?
			How many had jobs?
			Which was the occupation that was at WORK?

Functioning	Theme	Sub Theme	Specific Question in Census of Brazil 2010
			Type of Employment
			Monthly gross income
			In MAIN JOB, HOW MANY HOURS USUALLY WORKED PER WEEK
	Access to basic services	Access to basic services	Household features: Sanitation, Water Supply, Refuse Disposal, Energy Source
	State funded programmes	Provision of state grants targeted at poor households	RETIREMENT or PROVIDENT INSTITUTE PENSION OFFICER (FEDERAL, State or MUNICIPAL)
			BOLSA Familia SOCIAL PROGRAM or PROGRAM of ERADICATION OF CHILD LABOR-PETI
Being in a conducive home learning environment	Negative attitude to learning in the household, community	Negative Environment	How many people live in the household
	Child headed households	Child headed households	Characteristics of the resident - Age
		Migrant Caregivers	How long have you lived without interruption in this unit of the federation (State)?

Table 61: Questions from the Brazilian Census of 2010 related to functionings valued by learners

The Brazilian Census of 2010 was found to be strongly accurate, clear, comprehensive, consistent and possess integrity. For each of these factors the census achieved the perfect score of 1. The weakest factors were conciseness and data currency each achieving a score of 0.5 whilst all the data traceability factors were found to be non-applicable as no source system was relevant to the collection and collation of the census results. Perhaps the primary reason for the strong performance of the Census is the detail offered and attention paid to the metadata production. Whilst a perfect score was not achieved, the detail that was provided supported the perfect scores achieved in other data quality dimensions of an architectural nature. Other strong dimensions that were accessed include data accessibility and applicability with scores in the 0.6 to 0.7 range.

The strength and usefulness of the metadata was found in terms of all three related indicators specifically the content, context and structure of the metadata provided. The primary concern within the metadata related to the reporting of the physical structure of all the released files which may have been an oversight by the IBGE. In attempting to provide the data users with a range of table options of analysis of the broad survey, the IBGE released many tables on their website as part of the Census data release, however every table that was released was not accompanied by a supporting metadata codebook which detailed the physical structure of each field. Codebooks were not provided for the numerous small tables that were made available, thus in producing multiple data reports, the net effect actually hindered the data user’s understanding of what the data represented. Except for this particular issue, the Census performed well as data dictionaries, statistical formulae reporting, sampling techniques and findings reports were all well documented and provided sufficient detail and support to the data user. Whilst the metadata did comprehensively address the various tasks that were undertaken to achieve data quality, the metadata does not explicitly define and measure metadata. This has been found to be a consistent challenge across many datasets in each of the three countries.

In terms of conciseness and currency of the data collection, the following reasons were identified for the imperfect score that was achieved. Whilst data conciseness actually refers to the minimalist nature of data provided, it may be found that the granularity of data may become an issue as the data did not provide information for areas smaller than sub-districts. The limitation related to the currency of the information was simply due to the infrequency of a national census, which is conducted every 10 years (although this is a norm internationally). As the information is of such great value, users would prefer such data to provided more frequently.

The concerns that related to data applicability referred to a common concern across datasets which was the lack of information regarding the users’ data requirements. If the users’ data requirements were met, they were not discussed within the metadata. In terms of accessibility, the main concern, albeit limited in nature, pertains the lack of a data release schedule and the lack of database tools that could be provided to assist the data user in working with contents of the survey across the many variables on offer.

The dimensions of accuracy, clarity, comprehensiveness, consistency and integrity achieved perfect scores which entails that no focal issues raised were linked to any negative perspectives. Therefore, in terms of accuracy we found that the data coverage methods were adequate and the data was suitable for reporting. Clarity of data use refers to how terms are used and defined and if the reported data consistently referred to the same set of terms referred to the metadata. The data is judged to be fully comprehensive as the dataset contained all statistical units alluded to within the metadata and all the data rules could be tested against the actual data released. In terms of consistency, we find that the dataset reports on the same set of terms discussed within the metadata and the data was found to possess integrity as data constraints applied during data capture were all thoroughly documented.

The main issue we find with the Brazil Census is the mechanism the IBGE provides to the public to extract and download data. Without database tools, users find it difficult to work out which released spreadsheet is the most appropriate to use. The IBGE could investigate releasing the census using a single Microdata package or provide metadata for every spreadsheet that is released. The data quality scores that were achieved per dimension are presented in Figure 16.

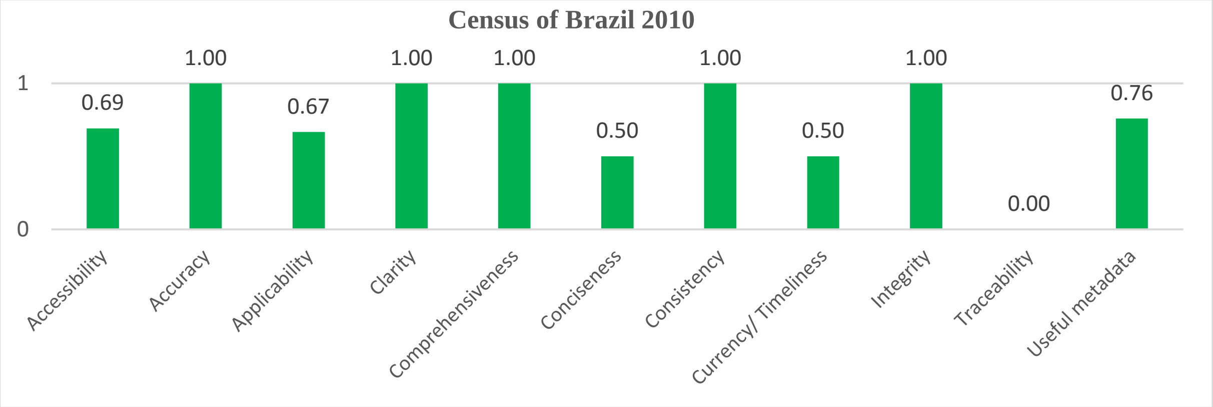


Figure 16: Data Quality Scores of the Brazil Census 2010

Brazilian Crime Statistics 2014

The Brazilian Crime Statistics are carried out the by the National Secretariat of Public Security or Secretaria Nacional de Segurança Pública (SENASP) in Portuguese. The system is based on the data collected by the various public security organisations of the Brazilian Federation. The system itself is called the National System of Public Security Information, Prison and Drug (English translation) which is focussed on standardising and organising the flow of crime data amongst the various public institutions. The work to date has focused on consolidating the disparate data collections to produce a singular data release describing the various incidents of crime across Brazil (Sistema Nacional de Informações de Segurança Pública (SINESP) n.d.), however the focus by this institute has placed no emphasis on producing detailed metadata which describes the processes followed and the standards adopted (if any exist) in setting up the system. Due to the absence of any metadata or supporting findings report released, the data collection instrument receives a very weak score of -7.1. The data is made available in individual reports per type of crime at a state level. Such data is useful in determining the level of crimes affecting households in particular communities and highlights the disparities such communities must contend with.

Functioning	**Theme**	**Sub Theme**	**Specific Data Points in the Crime Statistics**
Meaningful participation in school	Household and Community Environment	Crime, violence, substance abuse and other negative environmental factors	Types of Crime per area

Table 62: Fields from the Brazilian Crime Statistics related to functionings valued by learners

Looking specifically at the data quality assessment, each factor scored weakly apart from conciseness which scored 0 as the data included fields which were relevant. All of factors were weak primarily due to the lack of any metadata. Therefore, in relation to the question of whether the released metadata is useful, there is no metadata to consult to make any determinations of data quality. This single issue affects all the other dimensions whereby we find that that data comprehensiveness cannot be assessed due to no benchmark to compare against. Similarly, for data accuracy, the dimension can also not be assessed, however, the only positive focal issue in relation to accuracy is that SINASP is audited successfully which suggests there are some documented controls in place, although not released in accompaniment to this dataset.

The other notable concern raised in relation to the data is the frequency of the data releases. Crimes statistics are collected daily and released annually but these statistics are released after two years, thus reducing the timeliness of the data release. In terms of data consistency, the organisation should be commended on providing a consistent data release with all reports consistently referring to a common set of statistics and terms. Considering the mammoth task of consolidating reporting systems from multiple data sources across multiple states, SINASP has performed well to succinctly provide data online for public use.

Despite the lack of published metadata, the crime statistics provide the public very valuable insight into crimes as these registers are valuable tools which provide a rare set of information. However, with no published metadata, the user must assume the negative focus issues exist instead of the positive due to the lack of evidence. SINASP would benefit greatly if they partner with the likes of the IBGE in order to gain insight in how to adopt greater quality assurance practices when collecting such data.

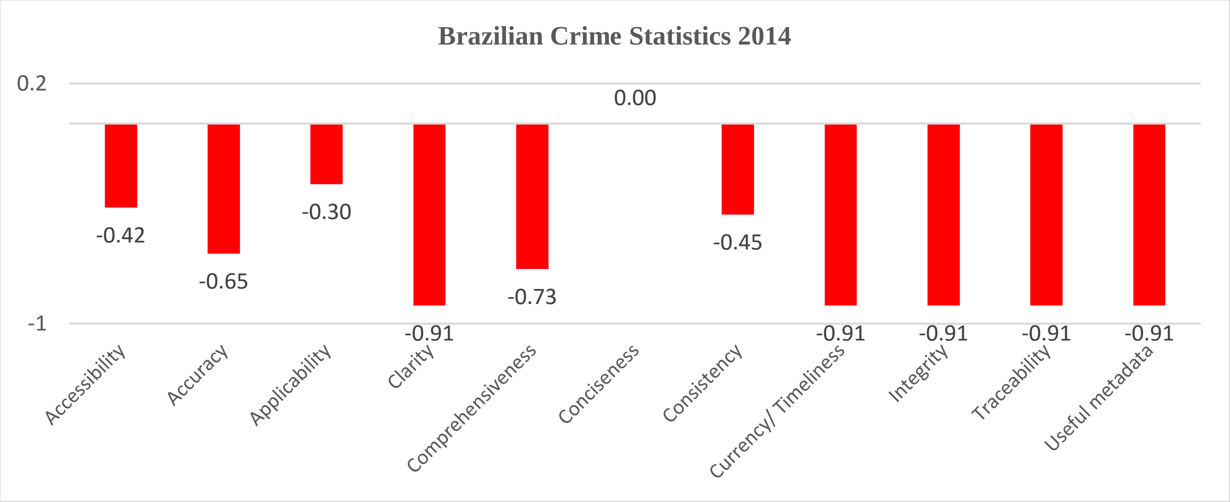


Table 63: Assessment of Brazilian Crime Statistics of 2014

b. Indian Data Collection Instruments

Within India, four data collection instruments were identified that provide data pertaining to the functionings identified in the literature regarding why learners drop out of school. The Census of India provides population and general socio-economic trend information, the Unified District Information System for Education (U-DISE) provides information collected by the educational organs of state, the Demographic and Health Survey provides trends about Health related matters in general about the country and the Crime Statistics provide trends on occurrences of crime across the various states and districts of India. Due to the vast spatial and geographic challenges of the country, collecting representative data is difficult in India. These challenges are echoed in the data quality scores attained by each data collection instrument which are discussed in the below table.

Data Provider	Collection Instrument	Score
Office of the Registrar General & Census Commissioner, India (ORGI)	Census of India 2011	0.68
National University of Educational Planning and Administration (NUEPA)	Unified District Information System for Education (U-DISE) 2014/15	1.04
DHS Program	Demographic and Health Survey 2005/6	5.26
Ministry of Statistics and Programme Implementation, National Crime Records Bureau, Ministry of Home Affairs	Indian Crime Statistics	-7.20

Table 64: Assessment of Indian Data Collection Instruments

Census of India 2011

The Census of India is conducted every ten years and is managed by the Office of the Registrar General & Census Commissioner, India (ORGI). The Census provides data on the financial security of the learners, the home learning environment and the physical health of the learners (Government of India Ministry of Home Affairs Office of the Registrar General & Census Commissioner n.d.). The detail and depth of the information makes it very attractive as a data resource, but the collection infrequency limits the currency of the information from the perspective of the data user. Furthermore, due to the challenges faced in organising the mammoth programme, various details needed by users in the metadata are neglected, and together with a restrictive data access policy the Census has attained the low data quality score of 1.43.

Functioning	**Theme**	**Sub Theme**	**Specific Data Points in the Census of India 2011**
Being financially secure	Financial Security	Socio-economic status of the learner, school and community	Primary Census Abstract Data for Slum (India & States/UTs - Town Level) (Excel Format)
		Low Household Income	Marginal Workers and Non-Workers Seeking/Available for Work Classified by Educational Level Age and Sex (States/UTs)
			Households by Size and Number of Members Seeking /Available for Work (India & States/UTs)
			Main workers, Marginal workers, Non-workers and those marginal workers, non-workers seeking/available for work classified by
			Non-workers by main activity, educational level and sex - 2011(India & States/UTs-District Level)
Being in a conducive home learning environment	Negative attitude to learning in the household, community	Education attainment of the head of household and associated community	Single Year Age Returns by Residence, Sex and Literacy Status (India & States/UTs)
			Educational Level by Age and Sex for Population Age 7 and above (Total, SC/ST) (India & States/UTs-District Level)
			Main Workers by Educational Level, Age and Sex - 2011(India/States/UTs/District)
		Negative Environment	Normal Households by Household Size (Total, SC/ST, City)
			Houseless Households by Household Size (Total, City)
	Child headed households	Child headed households	Households by marital Status, sex and age of the head of household (Total, City)
Being physically well	Menstruation and Female Maturation	Gender biased communities	Ever Married and Currently Married Population by Age at Marriage, Duration of Marriage and Educational Level -2011(India & States/UTs/District level)

Functioning	Theme	Sub Theme	Specific Data Points in the Census of India 2011
			Number of Women and ever Married Women by Present Age, Parity and Total Children Ever Born by Sex (India & States/UTs - District Level) (Total, SC/ST)

Table 65: Fields from the Indian Census 2011 related to functionings valued by learners

In review of each data quality dimension, the Census scores positively and negatively for each dimension. The negatives are largely offset by the positives resulting in scores of 0 for each dimension apart from data comprehensiveness and useful metadata which is slightly positive when all focal issues are summarised across the dimension. On closer inspection of the metadata usefulness dimension, the bulk of the negative focal issues related to the context of the metadata. In terms of the content of the metadata, the concepts, fields, tables are geographical boundaries are discussed in the supporting documentation. However, when looking the contextual issues within the metadata, it is found that the metadata does not describe in sufficient detail the data quality practices followed by the Office of the Registrar General and Census Commissioner. Although the metadata alludes to a pilot study that was set up to investigate the pressing concerns for the census, the metadata does not follow that with specifics about data quality management, statistical assessments or a discussion on how data quality concerns were addressed during collection. Furthermore, the metadata does not provide sufficient detail to allow the data user to understand whether the data collection comprehensively reached the targeted population. Understanding that reaching the entire population of India in a single survey is difficult, the published metadata does not provide any information regarding the response rate of collection. Furthermore, the questionnaire structure is not discussed and the metadata does not provide any references to international norms and standards which guide how the census is carried out. On the positive side, a detailed findings report does accompany the release of the data and the use of concepts are consistently applied throughout the documentations. In terms of the structural issues affecting the metadata, it is found that the physical structures of the census are documented however the software and hardware needs for accessing the information are not referred to explicitly.

In terms of data comprehensiveness, it was found that the data files provided a summary of all fields mentioned within the metadata. However, the data dissemination strategy that was followed entailed providing only static excel spreadsheets which limited the analysis as multivariate cross-tabulations were more difficult to perform across spreadsheets. In terms of the detail provided, each file described all fields that were mentioned within the metadata code book, the only limitation in use is the inaccessibility of analysing the database in its entirety. To access greater detail, users needed to find Census University Workstations which are located at a selection of universities across the country. At these workstations, detailed databases are installed allowing users the opportunity to more closely examine the census results. However, these installations were quite restrictive as one's results could only be extracted by physical printout with no digital downloads of the information allowed. If a user requires a particular cross-tabulation of data not provided in the static spreadsheets made available online, a request is needed to be submitted to the ORGI staff. Another accessibility limitation of the provided data is that for every published spreadsheet, was that documentation is not provided to assist the user in interpreting the results of the data and the purpose of the particular spreadsheet.

In terms of data accuracy, the primary weakness of the census is the lack of information provided regarding the assessments of data coverage and data suitability for reporting. With a lack of discussion in sampling techniques or imputation techniques the user is unable to make an assessment of the data quality achieved.

In terms of data clarity, applicability, conciseness and consistency, one finds the census data collection processes performed adequately well, as naming conventions are consistently applied, the information provided is highly useful, although it does not fully capture whether the user's needs have been met. These needs are not documented even though a pilot study is referred to in the documentation. With regards to the conciseness of the data, one notes that although each field referred to within the metadata is provided within the released static data files, the lack of flexibility severely inhibits the user's ability to analyse the pertinent information provided. In terms of data consistency, the development of the census questionnaires is noted in how the terms have evolved over time, which is documented. However, the user is not informed whether the terms follow any international standard. Often the language used follows from terminology specific to India alone, which also inhibits data comparability.

With regards to data currency and integrity, it is noted that the census website and documentation does not refer to a data release schedule but due to the long time line, this factor is not viewed negatively. In terms of data integrity, a major limitation of the metadata is that it does not discuss the data quality constraints applied by ORGI during data collection. In terms of the data transformations that are applied, whilst there are numerous spreadsheets provided, the bulk of the data released matches the finding reports provided the user some surety on the correctness of the data transformations applied.

With respect to the Indian Census of 2011, the major impediments found refer to the lack of detail within the metadata pertaining to the data quality processes such as response rates or imputation rates as well as the restrictive data dissemination policy. The lack of online data tools or the release of granular data downloads limits the public from performing detailed analysis. The only users that can perform such tasks are those close to universities which house census workstations. Even in these situations, data is not freely accessible as no data downloads from the workstation are permitted. The scores achieved are presented in Figure 17 below.

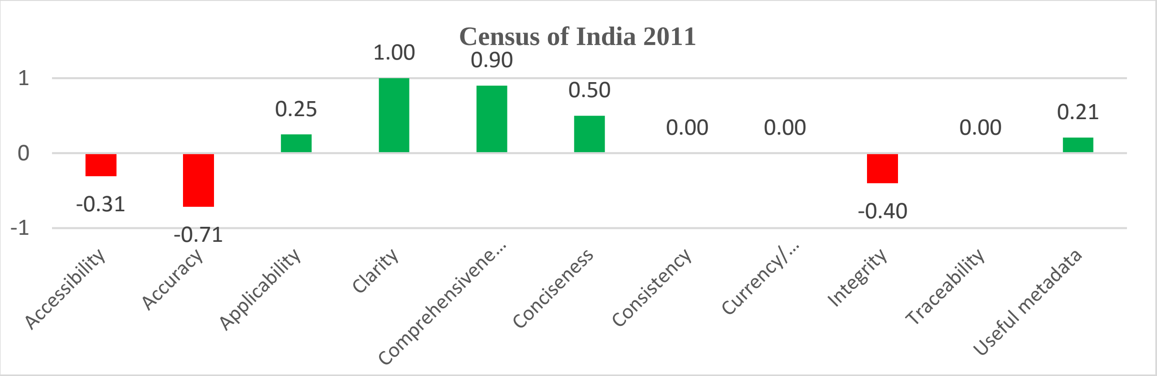


Figure 17: Data Quality Scores of the Brazil Census 2010

Unified District Information System for Education 2014/15

The Unified District Information System for Education as managed by the National University of Educational Planning and Administration provide statistical publications each year on the performance of the educational ministries in the country. The data collected is highly pertinent to the monitoring of schools in India and contains details such as learner access to schools, skills and training of educators, school feeding programmes, facilities available within each school as well as the literacy and numeracy levels of learners at different grades (National University of Educational Planning and Administration (NUEPA) n.d.). With such a vast array of information available, it is noted that the assessed data quality score for the data release is 1.04 highlighting that similarly as experienced with the Indian Census, there are a number of data quality concerns that persist.

Functioning	Theme	Sub Theme	Specific Data Points in the Census of India 2011
Being physically well	Malnutrition and Hunger	Access to feeding programmes	Schools Providing Mid-Day Meal (Government & Aided Schools)
			Schools where Mid-Day Meal is Provided and Prepared in School Premises (Government & Aided Schools)

Functioning	Theme	Sub Theme	Specific Data Points in the Census of India 2011
	Teenage Pregnancy	Late School Entry	Distribution of Enrolment by Age & Grade: All Areas
Traveling to school safely and conveniently	Distance to school	Available Access Road to school	Schools Approachable by All Weather Road
		Distance learner commutes to school	Schools in Rural Areas to All Schools
Being Taught in Suitable Infrastructure with necessary resources	Improve Physical Infrastructure	State of disrepair of the school	Distribution of Schools by Status of Building, All Areas: All Managements
			Distribution of Classrooms by Condition of Classrooms: All Managements
		Ensure that the schools are of a suitable size	Single-Classroom Schools
			Schools with Enrolment <= 50
			Number of Classrooms by School Category: All Government Managements
			Average Number of Classrooms by School Category: All Private Managements
			Distribution of Schools by Number of Classrooms: All Managements
			Student-Classroom Ratio, All Areas: All Managements 44
			Number of Classrooms by School Category
	Facilities and Resources	Available facilities within a school	Schools by available facilities (boundary wall, toilets, computers, drinking water, electricity, kitchen shed, playground)
Free Expression of Opinions	Conducive school learning environment	Appropriate number of teachers	Student-Classroom Ratio and Pupil-Teacher Ratio at Primary and Upper Primary Level
			Single-Teacher Schools
			Teachers by School Category
			Pupil-Teacher Ratio by School Category: All Areas: All Managements
Meaningful participation in school	Comparative academic strength of the learner	Understanding elementary level concepts	Literacy Rate by gender
	Comparative academic strength of the learner	Comprehension of classwork	Examination Results, All Areas: All Managements
	Teacher Quality	Pedagogical skills and knowledge for teaching	Percentage of Teachers by Professional Qualification: All Managements
			Teachers Received In-service Training
			Professionally Qualified Teachers by school category

Table 66: Fields from the Unified District Information System for Education 2014/15 related to functionings valued by learners

In terms of data quality, as found with other data collection instruments, the data quality concerns follow the level of detail discussed within the metadata. Firstly, with respect to the content covered in the metadata, the terms and concepts are discussed within the education planning handbook although these handbooks were last updated in 2003 and are not current. Together with this is that the individual surveys or data collections conducted by the Ministry are not referred to within the documentation and neither is every field and table that is provided, documented. With regards to the context of the metadata that is provided, the data quality practices are discussed to an extent where data sampling techniques are referred to, as followed by DISE, however how one measures data quality in the organisation is not discussed extensively. In terms of how data quality concerns are processed, it is noted that random sampling reports are produced to discourage individual school districts from performing incorrect data collections. With regards to the details of the data collection methodology, some insight is offered in the metadata which details how the ministry targets every school and how 5% of schools are surveyed to check data correctness. However, these discussions are not well explained and do not refer to individual questionnaires that are circulated to schools.

In terms of data clarity, comprehensiveness, conciseness and currency, the U-DISE released data performs well. In terms of clarity, all the terms and concepts used in data files refer to specific terms found in supporting documentation. With respect to data comprehensiveness, it was found that the released datasets provided information on every field combination that was provided, was also discussed within the metadata. With regards to conciseness, it was noted that all the data fields released were relevant however pertinent school level information was not released. In addition, it was noted that the organisation has no particular strategy for users to access such data. The currency of the data was also found to be a sore point as the data is released annually, but with a two-year delay from the point of collection. Furthermore, no time-frame is released to accompany the dataset informing the public of when to expect the data.

In terms of data applicability, accessibility, accuracy and integrity, various weaknesses are highlighted within the data collection process. In terms of accessibility, whilst the data is made freely available, the format of the data provided is limiting with all data tables provided in PDF format. Furthermore, detailed school level information is restricted with no school level information made available and no guidance offered to users on how to access more detailed data. In terms of data applicability, there was no mention of the user's requirements for using the data. On the positive side, it is noted that no technical details that were irrelevant to data users is included. In terms of accuracy, the dataset scored positively as a data discrepancy rate of 5% is referred to in the documentation and NUEPA is regularly audited, however on the negative side, the imputation techniques that are followed, are not discussed and the testing of only 5% of schools is low. Furthermore, where the 5% of schools are contacted, no effort is made to correct the information collected based on the tested data. Lastly in terms of data integrity, no references are made to ensure that any data constraints are applied at all when collecting the data and neither are the data transformations documented.

One of the primary concerns with the U-DISE data collection, apart from the metadata weaknesses is the limited release of data in usable formats. By providing multiple individual datasets in PDF format, the format is inflexible for data analysis and the multitude of reports makes it difficult for the organisation to provide supporting documentation regarding the use and purpose of each tabulation. Larger datasets with a singular piece of documentation describing the data collection effort as a whole would be more useful to the data user as well as be simpler to manage by NUEPA. The summary of the data quality dimensions is highlighted in Figure 18 below. Furthermore, it is recommended that greater emphasis is given to managing data quality constraints during data collection and thereafter documenting such practices within the metadata.

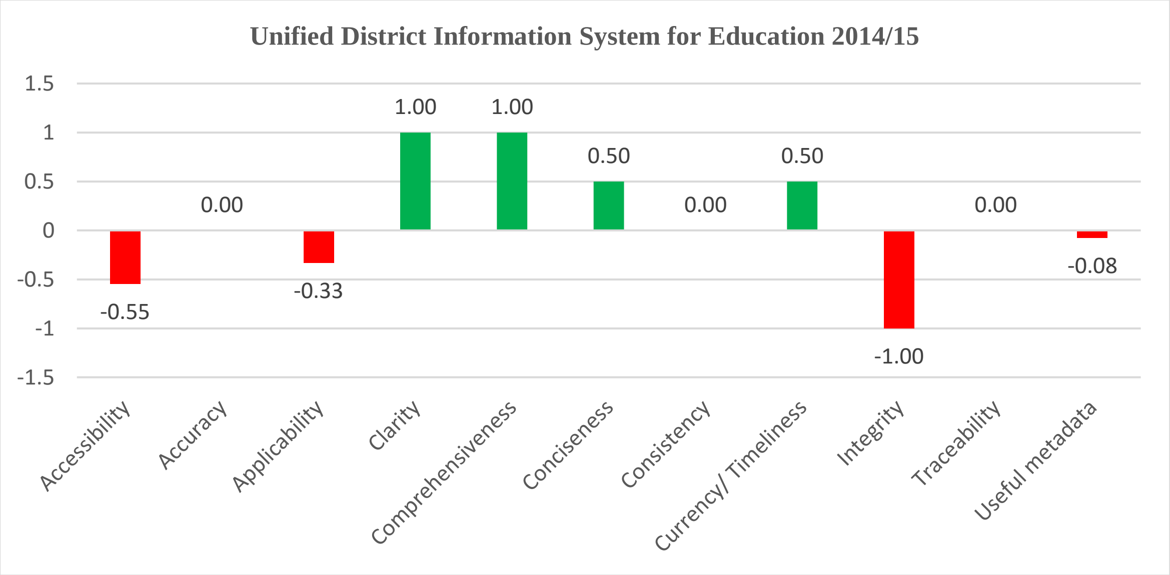


Figure 18: Data Quality Scores of the Unified District Information System for Education 2014/15

Crime Statistics of India

The Ministry of Statistics and Programme Implementation (MOSPI) together with the National Crime Records Bureau and the Indian Ministry of Home Affairs release on the MOSPI website and within the India in Figures publication statistics on the incidents of crime occurring across the country. The documentation released by MOSPI discusses the relationships held between MOSPI and the Crime Records Bureau as well as with the United Nations Statistical Division and the various categorisations of crime but the documentation is light on the data collection practices and provides little information regarding how data quality is protected by any of the data collectors (Government of India Ministry of Statistics and Programme Implementation 2005). Consequently, in terms of the data quality assessment the dataset scores weakly with a value of -7.20.

Functioning	Theme	Sub Theme	Specific Data Points in the release Crime Statistics
Meaningful participation in school	Household and Community Environment	Crime, violence, substance abuse and other negative environmental factors	Incidence of Cognizable Crime Under IPC, All India, State Wise and City Wise.
			Juveniles Delinquency IPC Cases All India and State Wise
			Juveniles Apprehended Under Cognizable Crime (IPC+SLL) All India and State Wise
			Juvenile Delinquency Under Special and Local Laws (SLL) (Cases Reported), All India and State Wise
			Juveniles Apprehended by Age Group and Sex (IPC& SLL), All India and State Wise,
			Educational and Family Background of Juveniles Arrested Under IPC, Special and Local Laws, All India and State Wise
			Distribution of Juveniles Arrested Under IPC and Special and Local Laws by Economic Set-Up and Recidivism, All India and State Wise
			Disposal of Juveniles Arrested (Under IPC & SLL Crimes), all India and State Wise
			Number of Persons Arrested Under Different IPC Crimes by Sex, all India and State Wise
			Motives of Murder and Culpable Homicide Not Amounting to Murder, All India and State Wise
			Actual Police Strength
			Number of Cognizable Crime Under IPC and Strength of Police Force

Table 67: Fields from the Crime Statistics released by MOSPI related to functionings valued by learners

In review of the data quality assessment, the weak score attained by the released Crime Statistics is in the main due to the lack of detail provided in the supporting documentation for the dataset. Not enough substance is provided to the data user for them to gain any sense of the content, context or structure of the datasets that are collected. A chapter from 2005 is made available on the MOSPI website regarding the data collection process and the organisation of the data in terms of categories consumed by the United Nations Statistical Division but no particular information regarding the specific fields and tables that are provided is discussed.

When reviewing the data quality dimensions, one finds there are few positive aspects to report on due to the lack of supporting documentation. Each dimension is assessed negatively, however the few positive points reflect that there are some data quality initiatives in place that should be promoted ad expanded. Firstly, in terms of data accuracy, the National Crime Bureau is audited annually. The contents of the audit reports need to be assessed to determine whether the data collection processes have also been duly evaluated as well. Secondly all the information is highly applicable to the data users, as the crime records provide pertinent information pertaining to the security of the learning and the stresses that learners are exposed to. Thirdly, in terms of consistency, the reporting format for the crime statistics has stayed constant over the past 10 years with few data changes. The last positive pertains to data anonymity which is maintained as the data provided is aggregated to state and municipal level.

In summary, the data provided by MOSPI is highly useful to the data users but not much effort is made to report on the data quality practices or to make the useful information more accessible. Greater documentation needs to be made available to help the data user make an assessment of the associated level of data quality to help build their level of trust when referring to such data. The specific scores per dimension are provided below in Figure 19.

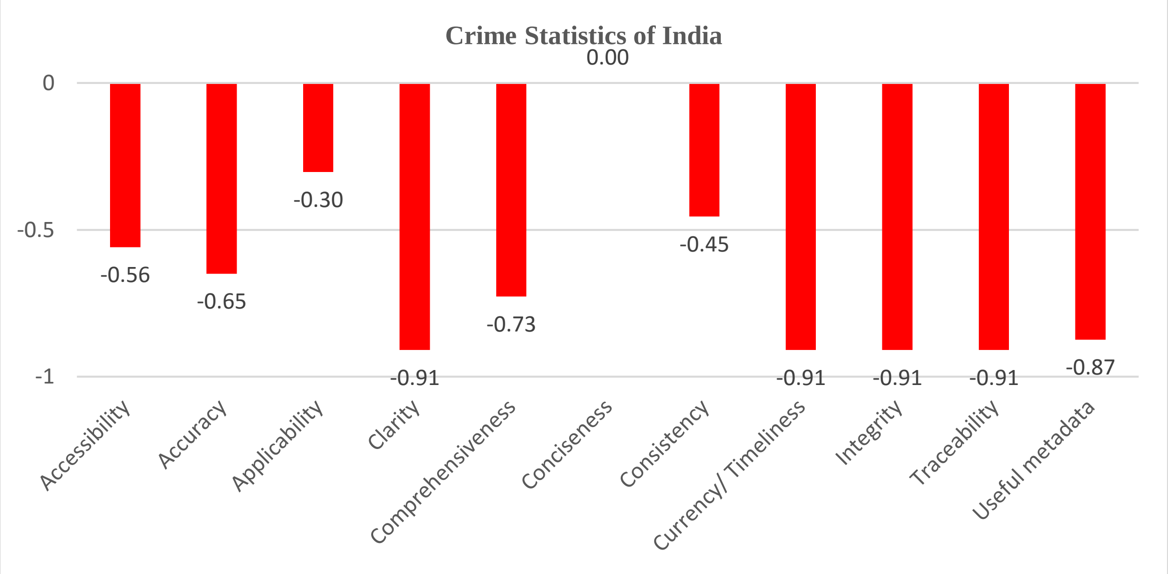


Figure 19: Data Quality Scores of the Indian Crime Statistics

Demographic and Health Survey 2005/6

The Demographic and Health Survey of 2005/6 for India is managed by the international Demographic and Health Survey Program which is funded by the United States Agency for International Development (USAID). The survey provides nationally representative data pertaining to trends on health and population matters. Due to the management by the DHS Program, the data is fundamentally comparable against other countries which partner in the survey (The DHS Program n.d.). The survey is particularly valuable as it provides rare information on teenage pregnancy and female maturation as well as information about the home and learning environment that learners face. Due to the many strengths of the data collection processes, the survey achieved a score of 5.26. The specific topics raised within the Indian survey are noted below in Table 68.

Functioning	**Theme**	**Sub Theme**	**Specific Data Points in the India: DHS**
Being physically well	Malnutrition and Hunger	Nutrition intake of learners	Micronutrients
			Vitamin A questions
			Nutrition - child feeding practices, vitamin supplementation, anthropometry, anaemia, salt iodization
	Menstruation and Female Maturation	Gender biased communities	Women's status
			Women's Empowerment - gender attitudes, women’s decision making power, education and employment of men vs. women
			Gender/Domestic Violence - history of domestic violence, frequency and consequences of violence
		Pregnant < 20 Years	Fertility and Fertility Preferences - total fertility rate, desired family size, marriage and sexual activity
	Teenage Pregnancy	Accessibility of contraception	Family Planning - knowledge and use of contraceptives
Being financially secure	Financial Security	Socio-economic status of the learner, school and community	Household and Respondent Characteristics - electricity, housing quality, possessions, education and school attendance, age, sex, employment
		Low Household Income	Wealth - division of households into 5 wealth quintiles to show relationship between wealth, population and health indicators
Being in a conducive home	Abuse in the household and community	Physical and verbal abuse	Domestic Violence (module) - prevalence of domestic violence and consequences of violence

Functioning	Theme	Sub Theme	Specific Data Points in the India: DHS
learning environment		Substance (drug and alcohol) abuse within the household and community	Alcohol consumption
	Negative attitude to learning in the household, community	Education attainment of the head of household and associated community	Education - literacy, attendance, highest level achieved

Table 68: Fields from India's DHS related to functionings valued by learners

In terms of data quality, the dataset performs strongly in terms of data clarity, comprehensiveness, conciseness and consistency, however the dataset's quality score could improve if certain items were included in the supporting documentation. This is reflected in the score for useful metadata, due to gaps that are identified in the documentation when describing data tables and the lack of metadata pertaining to the geographic granularity and structures. Although data on India is meant to be nationally representative, the metadata at no point confirms how granular the data is meant to be representative. In addition, an oversight within the metadata is the lack of definition or measurement of what data quality is. On the positive side, data sampling plans are provided and data collection scenarios are provided to help train field workers when conducting the survey, the organisation fails to explicitly define data quality and how it is measured and thereafter managed.

Other weakness of the dataset related to accessibility, applicability and currency. In terms of accessibility, the concerns that were highlighted related to the weaknesses discussed in the metadata such as the lack of documentation pertaining to the table structures and geographic structures. Furthermore, the documentation does not discuss the data dissemination strategy followed, therefore information such as the frequency and means of accessing the data is not discussed in detail and neither are the hardware and software requirements discussed. In terms of data applicability, the main concern is that the metadata does not describe whether the user's data requirements were assessed at all in order to determine what the public's priorities are in terms of health and population related information. In terms of data currency, the weakness that is found relates to the lack of information pertains to the infrequency of data collection and the publicising of a data release schedule despite the data collection that is conducted every 10 years.

In terms of the strengths of the survey, data clarity features strongly as the published metadata describes the concepts used in great detail and these follow the international standards adopted by the DHS Program. Data comprehensiveness is rated highly as the metadata suitably describes the statistical units of each field involved, and the dataset accurately conveys the details described within the metadata. Therefore, the structure and rules of the dataset are adequately described within the reported documentation. In terms of data conciseness, all fields provided included no superfluous information and were found to be to the point. Lastly with respect to data consistency, the dataset adequately represented the structural and semantic requirements that were stated within the rules of the metadata.

In terms of data accuracy, the dataset was scored highly but suffered on a few data elements which relate to reporting on the frame coverage of the survey in the supporting documentation. Therefore, it is unclear if the sample is truly representative of the entire country despite the statements that declare coverage is adequate. Furthermore, as the survey is conducted by an international body, it is audited as other national functions generally are within governments. On the positive side of data accuracy, the sampling frame follows international best practices and the confidence interval of the results was reported to be 95%.

In summary the weaknesses of the survey primarily relate to the lack of detail pertaining to the released tables and the lack of a clear data dissemination strategy. Due to these concerns, various elements were scored negatively despite the survey holding many positive traits which follow international best practices. The survey is claimed to be nationally representative, but due to a lack of published response rates, this cannot be confirmed. The scoring of the survey is found below in Figure 20.

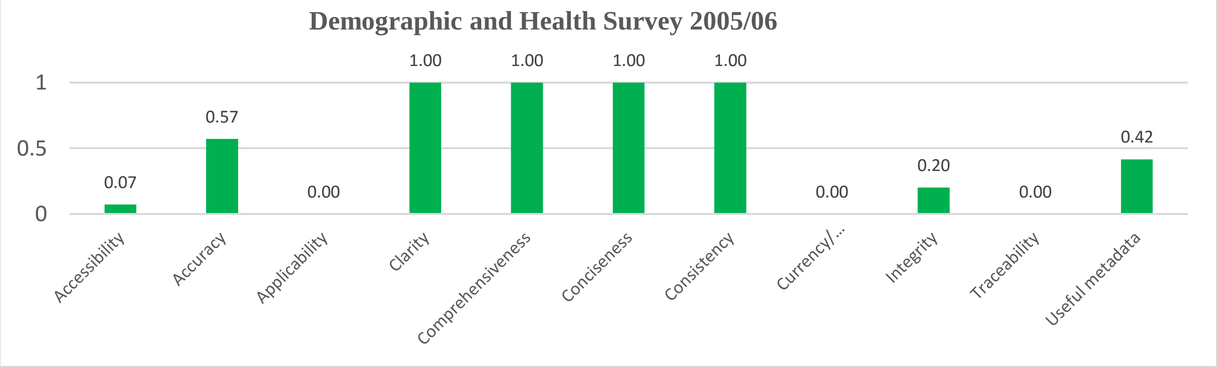


Figure 20: Data Quality Scores of the Demographic and Health Survey 2005/6

c. South African Data Collection Instruments

Within South Africa, thirteen data collection instruments were identified that provide data pertaining to the functionings identified in the literature regarding why learners drop out of school. These collection instruments are discussed in the below table. The data collection instruments are managed by government departments, statutory bodies and research institutes.

Data Provider	Collection Instrument	Score
Southern Africa Labour and Development Research Unit, University of Cape Town	National Income Dynamics Study Wave 1, 2, 3, 4	8.71
Statistics South Africa	Census of South Africa 2011	9.71
	Community Survey 2016	9.58
	General Household Survey 2015	8.70
South African Police Services	SA Crime Statistics	-7.29
South African Department of Basic Education	School Master List	-6.38
	Annual School Survey 2014 + Education Statistics at Glance Publication 2014	-6.90
	National Education Infrastructure Management System 2016	-6.90
	Annual National Assessment 2014	-6.71
SA Medical Research Council	Youth Risk Behaviour Survey 2011	0.97
	Demographic and Health Survey 20032	-4.90
Human Sciences Research Council	Trends in Mathematics and Science Study	6.50
	South African Social Attitudes Survey (SASAS) 2012	1.13

Table 69: Assessment of South African Data Collection Instruments

Department of Basic Education Released Statistics

The Department of Basic Education (DBE) is responsible for a variety of country wide collections using surveys and registers targeting all schools in the country. These data collection instruments include the School Master List which is used to manage all the schools in the country, the Annual School Survey which targets learners and educators in all schools, the National Education Infrastructure Management System which is used to track the state of school infrastructure across the country and the Annual National Assessments which target both learners and educators and seeks to identify the level of skills attained by learners and educators alike. For each of these data collections, it is noted that similar data quality concerns affect each collection instrument with the key findings published within the 'Education Statistics at a Glance' publication each year. Each of the DBE surveys and released dataset share similar data quality challenges whilst the datasets provide deep insight into the experiences at school, the trends in terms of learner progression and general information regarding the state of school infrastructure that learners need to contend with. Noticeably, each of the DBE datasets share a similar data quality scores, due to the similarities in their data collection processes and shared strategy in terms of data dissemination. The data quality score for each dataset ranges between -6.3 and -6.9, which further emphasises the similarity.

Functioning	**Theme**	**Sub Theme**	**Specific data points in DBE released dataset**
			Annual National Assessment
Meaningful participation in school	Comparative academic strength of the learner	Comprehension of classwork	Learners achievement levels for Grade 3,6,9 in Mathematics + Home Language, by province, in 2014.
			NEIMS
Being Taught in Suitable Infrastructure with necessary resources	Effects of Physical Infrastructure	School's access to basic services	Water, Electricity, Ablution Source Facilities
			Schools with Pits only and No Sanitation
Menstruation and Female Maturation	Menstruation and Female Maturation	Access to adequate sanitation in schools	Schools with Pits only and No Sanitation
			Schools by Ablution Source
Being Taught in Suitable Infrastructure with necessary resources	Facilities and Resources	Security of the school	Fencing and Security
	Facilities and Resources	Available facilities within a school	Sports, Communication, Library, Laboratory Source Facilities
			Annual School Survey
Being Taught in Suitable Infrastructure with necessary resources	Effects of Physical Infrastructure	Classrooms mixed by grade and subject area due to limited space	Schools with multi-grade classes, by province
		Classrooms mixed by grade and subject area due to limited space	Distribution of ordinary schools with multi-grade classes, by province and school size
Being financially secure	State funded programmes	Provision of free schooling	Number and percentage of ordinary public schools that do not charge school fees, by province
			Number of learners in ordinary schools, by province and funding type
		Provision of state grants targeted at poor households	Number of learners in ordinary schools receiving social grants, by province
Being physically well	Teenage Pregnancy	Late School Entry	Number of learners, by age and grade
		Pregnant <20 years	Number of learners in ordinary schools who fell pregnant, by province and grade
			Schools Master List
			Section 21 Classification

Functioning	Theme	Sub Theme	Specific data points in DBE released dataset
Being financially secure	State funded programmes	Cost of school fees and other resources	Quintile Allocation
		Provision of free schooling	No Fee School
	Financial Security	Socio-economic status of the learner, school and community	Urban Rural

Table 70: Pertinent fields and data points from DBE Data Collection Instruments

In review of the individual data quality scores per DBE dataset for each data quality dimension, one finds the score are generally similar which is due to common data collection and dissemination practices. Each factor apart from data conciseness is rated negatively. With the majority of dimensions scoring quite poorly, the roots of the data quality concerns begin with the lack of any released supporting metadata by the department. With the lack of metadata, the assessment of the content, context and structure of the metadata is impossible and therefore the assessment of each indicator and sub-element to the dataset is found to be negative. Importantly, there are no details communicated to the data user regarding the definitions of terms or any information regarding how the department defines and manages data quality. This lack of available documentation for the data user to peruse drives the poor scores for each following dimension.

In terms of data comprehensiveness, accuracy and clarity, one finds that each released dataset is impacted by the lack of supporting documentation. In terms of data comprehensiveness, the user is unable to assess whether all statistical units are provided in the data or if the organisation's data collection rules are applied correctly as the user is not informed of what the collection rules are. The School's Master List is published in its entirety on the DBE website, however, without any supporting documentation, the dataset cannot be assessed. With respect to data accuracy, the dimension also scores negatively with slight variances found in the Annual National Assessment as the DBE websites alludes to a pilot study which was conducted but without further details about the study or the findings. Apart from that point, each of the datasets provide no documentation to support an analysis of whether the data coverage methods were adequate or if the data is suitable for use. An important point to be considered, is that the DBE is audited annually although the results of the department's audit have not been examined in this study. In terms of data clarity, the lack of metadata again makes it impossible to denote whether the key concepts have been introduced and applied correctly throughout the dataset.

Data accessibility is found to be quite a challenge with respect to the DBE data collections. Whilst the School Master List is provided freely at school level on the DBE website, and some of the Annual School Survey data is published within the 'Education Statistics at a Glance' publication, the bulk of the DBE data is strongly protected and are not published online. Data requests to the department need to be made and responses are not always guaranteed. Data from these data collections is often referred to in documentation from the department and the data user is forced to consume '2nd hand analysis' of the data and make inferences on the data based on such information.

In terms of data applicability, the content of the data collections is found to be very useful when reviewing the broad state of education across South Africa, however the challenge facing the data users for each dataset is primarily the lack of documentation pertaining to the user's data needs. An option for the department could be to detail the department's particular motivation for carrying out the data collection in the manner that was followed. However, no documentation is provided to the data users to inform the rationale behind the data collection or if the public was consulted.

With respect to data reporting consistency, the lack of metadata once again makes it impossible to assess whether the data collection methodology pays attention to the semantic concerns such as data rule or structural consistency in terms of the table structures and fields used. In review of the data tables that are referred to in the 'Education Statistics at a Glance' publication, the tables used seem to reflect a consistent approach. However, this assessment is only superficial and therefore does not factor more strongly in the assessment.

In terms of data currency, integrity and traceability, similar concerns emerge related to the lack of supporting documentation. In terms of data currency, the findings reports are generally released over a year from the date of collection and no timeframe is offered to the public concerning when to expect the data release. In terms of data integrity, the data quality constraints applied during the data collection are not discussed and therefore cannot be assessed. Lastly the lack of metadata also denies the user the opportunity to assess the traceability of the reported data. If there are multiple source systems involved in capturing education data, they are not documented and therefore the dimension cannot be assessed for each DBE data collection.

In summary, the DBE may actually be conducting various data quality controls when collecting this data, but their efforts are not exposed to the public and therefore the user is forced to assume no quality processes are in place. Thus the DBE data collections are assumed to be unreliable due to the lack of information provided to the public. The scores per dimension for each DBE data collection is provided below in Figure 21.

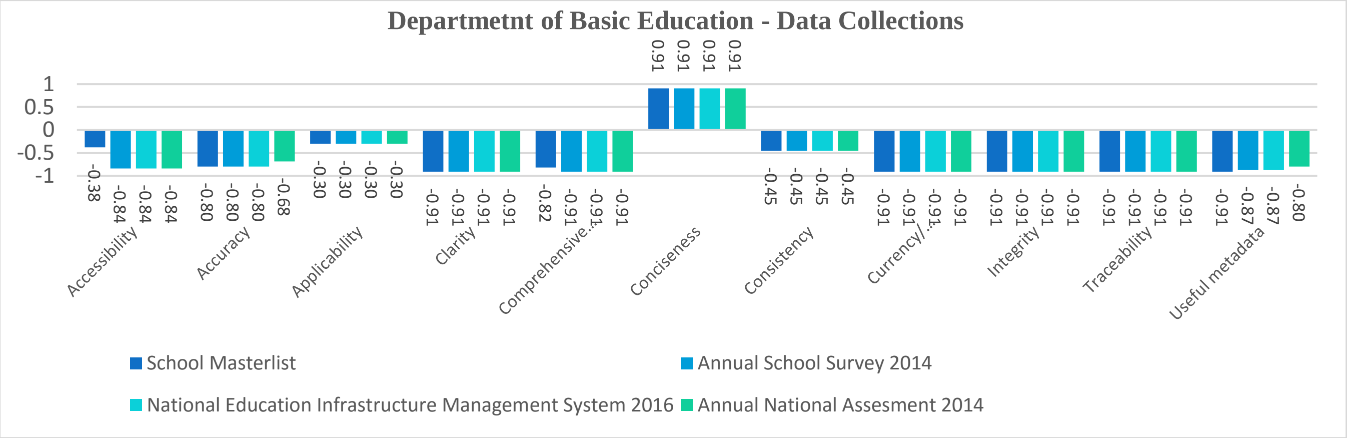


Figure 21: Data Quality Scores of Education Statistics released by the DBE

Statistics South Africa Data Collections

Statistics South Africa produces three surveys which are relevant in the assessment of learner dropouts. This includes Census 2011, the recently released Community Survey 2016 and the General Household Survey 2014. The surveys produced by Statistics South Africa predominantly expose information pertaining to population demographics. Using the person age variable, the demographics released can be filtered to the school going age to determine the trends affecting the learner population. The functionings that are related to these data collections are being financially secure, being in a conducive home learning environment, being mentally well, being physically well, traveling to school safely and conveniently and the learner's meaningful participation in school. The specific questions asked in relation to these functionings are described in Table 71.

Functioning	**Theme**	**Sub-Theme**	**Question**
			Census 2011
Being financially secure	Access to basic services	Access to basic services	Access to piped water, Energy Source, Toilet Facilities, Refuse Removal
			Source of Water
	Financial Security	Low Household Income	Employment Status of Head of Household
		Socio-economic status of the	Dwelling Type
			Construction Material of the Main Dwelling

Functioning	Theme	Sub-Theme	Question
		learner, school and community	
Being in a conducive home learning environment	Child Headed Households	Child headed households	Mother/Father Alive
	Negative attitude to learning in the household, community	Education attainment of the head of household and associated community	Qualification of the Head of the Household
			School attendance
			Level of Education
Being Mentally Well	Emotional factors	Stress	Death Occurred
Being physically well	Teenage Pregnancy	Pregnant <20 years	Age at first Birth
Meaningful participation in school	Comparative academic strength of the learner	Understanding elementary level concepts	Literacy
			Community Survey 2016
Being financially secure	Access to basic services	Access to basic services	Housing and access to basic services
			Access and quality on service delivery (water and sanitation, energy)
Being physically well	Malnutrition and Hunger	The effect of extreme poverty and hunger	Ran out of money to buy Food in past 12 months
			Households who skipped a meal in the past 12 months,
	Teenage Pregnancy	Pregnant <20 years	Children ever born
Meaningful participation in school	Household and Community Environmen t	Crime, violence, substance abuse and other negative environmental factors	Households' feelings of safety when walking alone
			Households' experience of crime
			Types of crime experienced by households
			General Household Survey 2014
Being financially secure	Access to basic services	Access to basic services	Type of dwelling of households, by main source of energy
			Main source of water for households, by province
			Households without water in the dwelling or on site, by the distance household members have to travel to reach the nearest water source, and population group of the household head
			Main source of energy used by households, by province
			Sanitation facility used by households, by province
			Refuse removal

Functioning	Theme	Sub-Theme	Question
	Financial Security	Cost of school fees and other resources	Population aged 5 years and older attending an educational institution, by annual tuition fee and type of institution
		Low Household Income	Sources of income for households, by province
	State funded programmes	Provision of state grants targeted at poor households	Population that received social grants, relief assistance or social relief, by population group, sex and province
Being in a conducive home learning environment	Negative attitude to learning in the household, community	Education attainment of the head of household and associated community	Population aged 20 years and older, by highest level of education and province
Meaningful participation in school	Comparative academic strength of the learner	Understanding elementary level concepts	Population aged 15 years and older with a level of education lower than Grade 7, who have some, a lot of difficulty or are unable to do basic literacy activities by sex and province
			Population aged 15 years and older with a level of education lower than Grade 7, who have some, a lot of difficulty or are unable to do basic literacy activities, by population group and sex
			Population aged 15 years and older with a level of education lower than Grade 7, by literacy skills and age group
Traveling to school safely and conveniently	Distance to school	Distance, time learner commutes to school	Distance travelled to get to the nearest minibus taxi/sedan taxi/bakkie taxi, bus and train, by population group of the household head

Table 71: Pertinent fields and data points from Statistics South Africa Data Collection Instruments

The three Statistics South Africa data collections have performed the strongest in terms of the data quality assessment that each dataset underwent. the Census of 2011, the Community Survey of 2016 and the General Household Survey (GHS) of 2014 respectively scored values of 9.7, 9.58 and 9.17 out of a possible 10 points. To achieve these particularly high scores, each survey scored strongly across each dimension where the assessment was applicable. Where the Census and GHS are from 2011 and 2014 respectively, the Community Survey data collection was only recently completed at the time of analysis and therefore the detailed release of data was not available.

The strength of the various dimensions is due to the strength of the reported metadata which accompanies the surveys. Statistics South Africa performs particularly strongly in this regard, as the organisation follows the SASQAF which is used in various components of the PDQAF. Therefore, in terms of the metadata's conveyance of the content, context and structure of the metadata, the documentation for each of the surveys adequately represents each of the requirement elements and indicators. It was further found that the concepts and terms used are well detailed and are accompanied by data registers which provide information about the relevant tables and fields made available per survey. In terms of the metadata's description of the context to which the dataset is situated within, a weakness across the datasets is the lack of measurements provided by Statistics South Africa on the level of data quality attained, although the metadata does clearly articulate the position the organisation takes in applying the SASQAF principles.

The strength of the Statistics South Africa dataset compared to the likes of the Brazilian Census or the National Income and Dynamics Study is the emphasis within the metadata regarding the importance of defining data quality using the SASQAF principles. Whilst in the other data collections that perform strongly, data quality is considered but does not feature as centrally when compared to the recognition that Statistics South Africa datasets placed on the application of the SASQAF in all aspects of their data collection processes. With respect to the data structures, the Census and GHS perform well in the manner the documentation describes the makeup of the individual tables and fields, however as the Community Survey has not been released as yet, the detailed data is not available for download and therefore supporting documentation of the individual table structures is not released and therefore cannot be assessed. A negative focal issue in relation the GHS, is the lack of a released handbook which details how the data user would peruse and operate the database tools used to provide GHS data, as is provided with the Census 2011.

In terms of data comprehensiveness, both the Census and GHS provide a full range of information which is described in terms of the statistical units that make up each particular field and the data rules can be tested in terms of the data output of both surveys. As the Community Survey has only recently been released, the survey is not released in detail and therefore cannot be assessed as the others are. In terms of data accuracy, the data coverage techniques are all assessed to perform well as the enumeration areas are well discussed, as well as are details regarding the imputation techniques and the response rates achieved documented and noted to be within international standards. In terms of the suitability of using the data for reporting, the results of the surveys were found to correspond with the trends found in the Pilot and Post Enumeration survey, in the case of the Census and the Community Survey. However, for the GHS, the data trends are compared against previous versions of the GHS itself instead of another external comparative data source. With respect to the maintenance procedures that are applied, the sampling frame has been found to be regularly updated for each survey, the SASQAF has been found to be fully adopted and audits are regularly conducted.

With respect to data clarity, applicability, conciseness and consistency each of the datasets are scored strongly due to the detail provided in the metadata which supports the assessment in terms of the identified elements and indicators of these dimensions. Firstly, with respect to data clarity it is noted that identified concepts within the metadata are referred to consistently without changing interpretations. In addition, these terms are identified and agreed to by an international council of experts convened by Statistics South Africa. In terms of data applicability, the variables within the three surveys are found relevant as the metadata actually refers to the processes of public consultation that are instituted or where subject area specialists are consulted to draft the terms for the particular data collection. However, following this public consultation, the metadata does not describe whether the public is satisfied with the final product produced, which highlights a gap in the Statistics South Africa documentation process. In terms of data conciseness, all the fields are judged to be relevant without the inclusion of any superfluous details. In terms of granularity, as the Community Survey is not published assessing the detail is not possible at this stage. Lastly, in terms of data consistency, all elements are scored positively for each survey as it was found that the questionnaire constructions generally followed a common methodology, the supporting documentation detailed where changes to particular questions were made in comparison to previous survey iterations and the data values were reported on consistently across the three surveys.

The greatest variances in data quality scores are found in regards to data accessibility. Firstly, many of the data quality elements are found not applicable when describing the Community Survey as the data has not yet been released. The weakness which are highlighted predominantly affect the GHS where the metadata was found lacking as it does not provide details about the GHS data dissemination strategy to reach the public, and related to the lack of a manual or handbook to inform the data user how to negotiate the available online database tools that are made available. Such documentation is made available for the Census and therefore should be similarly produced for the GHS publication. This oversight weakens the GHS data quality score in terms of accessibility.

With respect to the remaining data quality dimensions of currency, integrity and traceability, it is found that apart from the assessment of the Community Survey in terms of currency, all three surveys perform strongly as data is timeously released and in terms of integrity, the metadata provides in detail the steps taken to apply the data quality constraints adopted by the organisation and which are adopted by the field workers. With respect to traceability, the dimension is not applicable as the dataset is not dependent on any supporting system.

In review of the surveys conducted by Statistics South Africa, it is noted that the degree to which the SASQAF is implemented and discussed in detail in the published metadata is an example to the other members of BRICS as their surveys emphasize data quality during data collection and production. Ensuring these processes are well documented provides the data user a sense of surety and trust in the organisation that the data is robust and reliable. The scores attained by Statistics South Africa are discussed below in Figure 22.

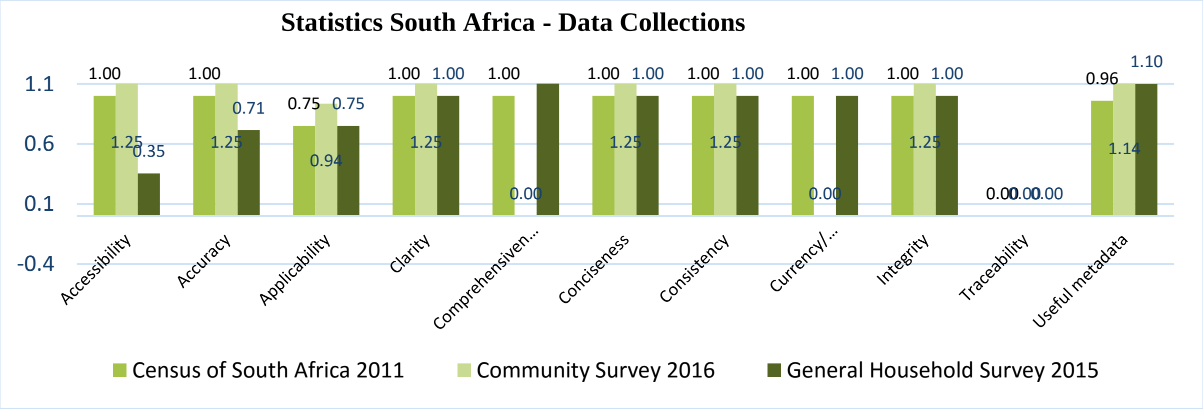


Figure 22: Data Quality Scores of the Census of South Africa 2011

National Income Dynamics Study Wave 1, 2, 3, 4

The National Income Dynamics Study is a panel study conducted across households in South Africa with a sample that is nationally representative. Four waves of the survey have been conducted since its inception in 2008. As the study follows the panel study methodology, all efforts are made to contact the initial 28000 respondents that were identified in the initial survey. Each wave targets these individuals and tracks their changes over time. Therefore, the survey is costly to manage, but produces very valuable and rare data insights across a range of subject areas such as poverty, well-being, child-headed households and notably trends affecting learners in schools. The study is managed by the Southern Africa Labour and Development Research Unit (SALDRU) based at the University of Cape Town (Southern Africa Labour and Development Research Unit n.d.). The identified relevant questions per identified functioning are described below in Table 72.

Functioning	Theme	Sub-Theme	Question
	Financial Security		Amount spent on allowances and other school related expenses

Functioning	Theme	Sub-Theme	Question
Being financially secure		Cost of school fees and other resources	Amount spent on Books and Stationery
			Amount spent on other school related expenses
			Amount spent on School fees
			Amount spent on Transport to school
			Amount spent on Uniform
			Did someone pay for child's educational expenses
		Low Household Income	Are you currently being paid a regular wage/salary;
	State funded programmes	Provision of free schooling	Is this a No Fees School?
		Provision of state grants targeted at poor households	Someone receives a social grant for the child?
Being in a conducive home learning environment	Child Headed Households	Child headed households	Biological Father is alive?
			Biological Mother is alive?
			Death of Mother before respondent was 5 years
			Father's Highest school grade
		Migrant caregivers	How often does the Father see the child?
	Negative attitude to learning in the household, community	Education attainment of the head of household and associated community	Father completed higher education?
			Mother's Highest school grade
			Mother's level of Higher Education
Being physically well	Teenage Pregnancy	Pregnant <20 years	Currently Pregnant?
Being Taught in Suitable Infrastructure with necessary resources	Improve Physical Infrastructure	Ensure that the schools are of a suitable size	Approximate number of learners in classroom byte
			Approximately how many learners are in the child's
			Number of learners in child's classroom
Meaningful participation in school	Comparative academic strength of the learner	Comprehension of classwork	Has the child ever repeated a grade
			Number of times child failed grade
			Respondent repeated grade 'x'
			The child failed grade
			The child failed grade
	Household and Community Environment	Learner’s motivation for school work	Number of days’ child was absent from school
Traveling to school safely	Distance to school	Distance, time learner	Approximate travel time to reach school from home
			Time taken to reach school

Functioning	Theme	Sub-Theme	Question
and conveniently		commutes to school	
	Provision of transportation	Available transportation services	What is the usual mode of transport to school
		Cost of transportation	Amount spent on Transport to school long

Table 72: Pertinent fields and data points from the NIDS Panel Study

The four waves of the NIDS panel study achieved a data quality score of 8.71 which underpins the great care that has been taken to produce high quality data outputs that the public can trust when reporting on national data trends. The dataset scores highly except in the inapplicable dimension of data traceability. The strong data quality assessment emanates from the detail provided within the NIDS supporting documentation. The metadata provides a high level of detail describing the content, context and structure of the published dataset. In terms of the content, definitions are clearly presented in an accompanying codebook for each wave and released data file, which includes well-defined concepts. In terms of the context of the dataset, the data quality practices that are applied are reported within the technical documentation which outlines the various processes followed by SALDRU. Furthermore, the documented response rate achieved by the survey is above 90% for each of the waves highlighting that data quality measures are recorded by SALDRU.

In terms of metadata comprehensiveness, the statistical checks followed by SALDRU are documented together with information pertaining to the sampling frame that was followed, the details pertaining to the questionnaire structure and the scope of the particular data collections. A minor weakness that was highlighted, was the documentation alluding to Statistics South Africa practices that were followed, but not clearly documenting the reliance on which practices of the SASQAF were adopted. Lastly in relation to the structure of the dataset, it is found that the individual data files are well documented and the handbooks provided, clearly inform the user how to access the various datasets provided.

Following the thorough provision of metadata, very few weaknesses emerge when reviewing the data quality assessment of the individual dimensions. Data comprehensiveness scores highly as the released data agrees with the content of the metadata, the data is understandable (clarity) as the metadata describes the underlying concepts well and the data is consistently reported as the metadata closely corresponds with the changes made to the metadata. The data is current as the survey is conducted every 2 years and is released timeously as per a data release schedule. Data accessibility scores highly due to the detailed documents that describe to the user how to access the data. Data access control is implemented, but SALDRU is found to be responsive to data queries supporting user data access. The data is also provided in multiple formats with supporting documentation guiding users how to work with such software. Lastly data integrity scores strongly as the metadata provides details regarding how the data quality constraints are applied and field workers are guided in how to address data quality breaches. Furthermore, before the data is released, the data is reviewed by a faculty ethics committee, in line with university research policies.

The weaknesses in the data collection approach are few but primarily relate to the data accuracy dimension. Whilst other surveys discuss a corresponding data source that can be used to test the results such as a pilot or post enumeration survey, due to the nature of a panel study, no such process is either followed or reported on. In addition, as the sample is only representative at a national level, geographic granularity is not necessary. Whilst greater granularity is not part of the terms of the project, greater detail in the survey would be appreciated by data users. Lastly, a minor weakness in relation to data applicability, is the concern related to the public opinion of the value of the data. Public perceptions of the data are not discussed within the documentation.

In summary, the NIDS survey is promising dataset neither produced by a governmental department or the national statistics provider. Following guidance from Statistics South African and other leading data bodies, SALDRU has produced a high quality dataset producing deep and valuable insights not provided by other data collection processes in the country. The individual scores attained per dimension are discussed in Figure 23.

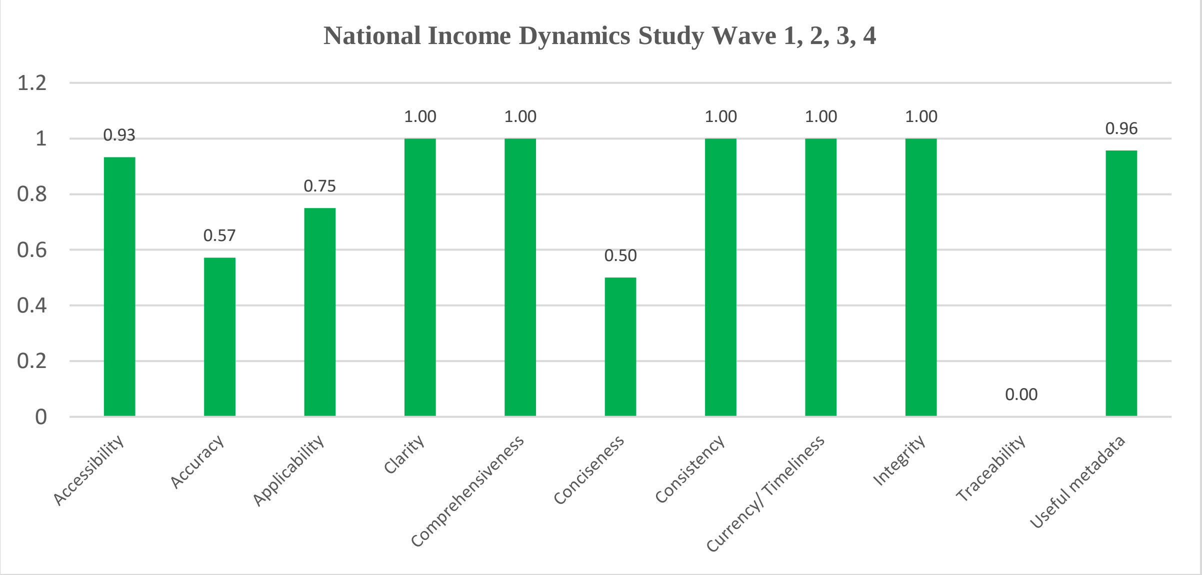


Figure 23: Data Quality Scores of the National Income Dynamics Study Wave 1, 2, 3, 4

South African Social Attitudes Survey 2012

The South African Social Attitudes Survey (SASAS) conducted in 2012 was managed by the Human Sciences Research Council (HSRC). The survey is conducted annually and is based on a provincially representative sample of 3500 to 7000 individuals aged 16 and older. The survey is used to monitor a set of general demographic, behavioural and attitudinal related questions and are supplemented with additional modules each year based on changing requirements based on pertinent current affairs at the time. The survey is used as a barometer to test the public's opinion on social, economic and political concerns (Human Sciences Research Council n.d.). Within the 2012 survey, questions that were relevant to school dropouts were identified which linked to functionings of the learner's financial security, their home learning environment, mental and physical well-being, free expression of opinion and the learner's ability to meaningfully participate in school. The questions identified related to the functions are described in Table 73.

Functioning	**Theme**	**Sub-Theme**	**Question**
Being financially secure	Access to basic services	Access to basic services	Do you have access to electricity in your household?
			What is the most often used source of drinking water by this household?
			What type of toilet facility is available for this household?
	Financial Security	Low Household Income	Before I buy something I carefully consider whether I can afford it
			How does your household income compare with other households in your village /neighbourhood?
			PERSONAL TOTAL MONTHLY INCOME before tax and other deduction
	State funded programmes	Provision of state grants targeted at poor households	Do you or anyone in this household receive any of the following Welfare grants?
Being in a conducive home learning environment	Negative attitude to learning in the household, community	Personal and Community's opinion of education value	From what you know or have heard, do you think school-leavers are better qualified or worse qualified nowadays than they were 10 years ago?
			How well do you think public secondary schools in South Africa nowadays prepare young people for work?
			How well do you think public secondary schools in South Africa nowadays teach young people basic skills such as reading, writing and maths?
Being Mentally Well	Emotional factors	Stress	Do you think that life will improve, stay the same or get worse in the next 5 years for people like you?

Functioning	Theme	Sub-Theme	Question
			How satisfied are you with your standard of living
			In the last 5 years, has life improved, stayed the same or gotten worse for people like you?
			Taking all things together, how satisfied are you with your life as a whole these days?
			Thinking about your own life and personal circumstances, how satisfied are you with your life as a whole?
Being physically well	Malnutrition and Hunger	The effect of extreme poverty and hunger	To what extent was the amount of food your household had over the past month less than adequate, just adequate or more than adequate for your household's needs?
	Menstruation and Female Maturation	Gender biased communities	Girls and boys should be educated separately
Free Expression of Opinions	Conducive school learning environment	Undisciplined learners	How well do you think public secondary schools in South Africa nowadays Instil discipline among young people
			respect teachers for their dedicated service to children and the community.
Meaningful participation in school	Discrimination against some learners	Discrimination in the household, school and community	Children of different religions, or of no religion, should be educated separately
			How often do you personally feel racially discriminated against?
			Where has this racial discrimination happened to you most recently?
	Household and Community Environment	Crime, violence, substance abuse and other negative environmental factors	Have you or a member of your household been the victim of a burglary or assault in the last five years?
			How often do you worry about becoming a victim of violent crime?
			How safe or unsafe do you (or would you) feel walking alone in this area after dark?
			How safe or unsafe do you (or would you) feel walking alone in this area during the day?
			How safe or unsafe do you feel personally on most days?
			How satisfied are you with how safe you feel?

Table 73: Pertinent fields and data points from the SASAS 2012

In review of the achieved data quality score of the SASAS 2012, the survey's score reflects a mixture of positive and negative reviews for the data quality dimensions. The overall quality score of 1.13 corresponds strongly with the positive and negative assessment of the usefulness of the metadata provided to the public by the HSRC. Whilst the HSRC has made attempts at providing supporting documentation to the public, various short-comings are exposed in terms of the context and content of the metadata provided. These weaknesses thereafter affect dimensions such as data integrity, consistency, applicability and accuracy in particular and this weakens other data quality assessments such as accessibility and conciseness. Furthermore, the HSRC data policy of embargoing the release of the most current iterations of surveys by two years severely impacts the currency and relevance of the data release.

On closer examination of the useful metadata dimension, the major concerns found under the context of the metadata, relates to the lack of discussion within the released documentation pertaining to the data quality assurance processes that are followed by the HSRC. Therefore, the data user is unable to determine how the HSRC defines data quality, measures data quality, addresses data quality concerns or performs their data collections. This lack of detail severely impacts the data integrity dimension where the data quality constraints need to be explicitly referred to in the documentation by including a discussion of how the HSRC manages their data rules. The limitations of the metadata's content also pertain to the manner in which concepts and terms are presented to the data user. A codebook of variables of the survey is made available, however the documentation does not express the context in which the concepts and fields are applied. In addition, the published data excludes a user guide to assist the data user in how to use the provided data.

The assessment of data accuracy is also undermined due to the lack of information pertaining to the data sampling techniques employed and the various tests that need to be applied to ensure the sample is representative. Although the survey is long running and is informed by subject area specialists, the metadata excludes key details required to assess accuracy. Furthermore, the metadata does not mention any comparative data source used to test the results of the survey or whether any data quality checks are included during data collection efforts. With the absence of such details, the user is forced to form a negative perception of the data produced.

Data consistency is also rated negatively as the metadata provided excludes information pertaining to the adoption of any international standards used when setting up the questionnaire. The consistency limitation is also found in the lack of details presented to the data user regarding the changes in structure to the questionnaire that may or may not occur after each annual iteration of the survey. On the positive side, the structure of data files provided are well explained with details regarding the software requirements to use the available datasets. In addition, a findings report of the available data is provided, highlighting the key trends found in the data. This mixture of positive and negative trends is summarised in the score of 0.09 for the useful metadata dimension.

Other negatively assessed dimensions are data applicability and data currency. The concerns pertaining to data applicability are the lack of public consultation involved in defining the terms of the survey. Whilst the survey is structured by subject area specialists, their priorities for the survey are either not tested through public consultation or if such processes occur, it is not documented for the benefit of the data user. The challenge related to data currency simply follows from the HSRC embargo against all recently conducted surveys. All surveys are not released to the public for two years after release. Such a practice is severely restrictive for an opinion based survey, as opinions can change dramatically over the course of two years.

The positive aspects pertaining to the data quality of the survey is found when assessing data clarity, comprehensiveness, conciseness and accessibility. The strength in terms of data clarity follows from providing a codebook for each data file and thereafter using consistent terminology in all published aspects related to the dataset. Data comprehensiveness scores strongly as the details provided within the dataset corresponds with the details itemised in the published metadata. All statistical units related to each question are provided and the data rules discussed within the questionnaire seem to be followed such that mandatory and inapplicable values are appropriately recorded within the dataset. However, the limitation of the released metadata which is highlighted in this particular dimension is the lack of detail about the data rules provided within the documentation. The conciseness of the data is reflected in the lack of irrelevant fields included in the published dataset. The positive data accessibility rating is based on the clear instructions provided on the HSRC website about downloading the available data and the many data formats that are made available to the data user to access the data. The concerns relating to data accessibility pertain the lack of public consultation in respect of their preferred means of accessing data.

In summary, the great value of the HSRC data collection is undermined by the lack of detailed metadata pertaining to the data collection processes followed and the documentation of the data quality practices employed. Furthermore, the HSRC data embargo detracts from the currency of the annual data collection, as each annual collection is released at a point when the data may no longer be relevant. The HSRC however can be commended in terms of their regular collection of data, the accessibility of data products released to the public and the comprehensiveness and clarity of the data that is made available freely each year.

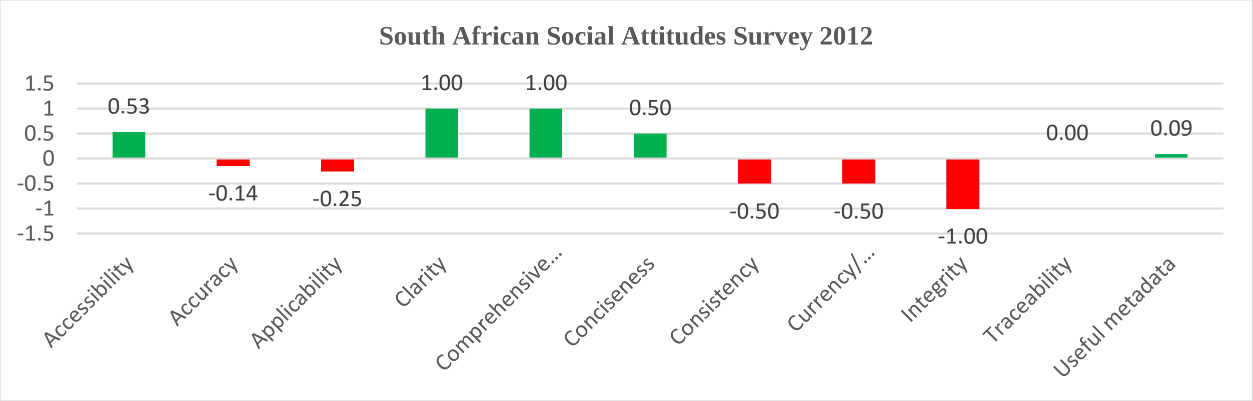


Figure 24: Data Quality Scores of the South African Social Attitudes Survey 2012

Trends in Mathematics and Science Study 2011

The Trends in Mathematics and Science Study (TIMSS) is also conducted by the HSRC for South Africa following the international guidelines specified by the International Association for the Evaluation of Educational Achievement. The survey was conducted in South African in 2003 and 2011. In 2011, the study was conducted in 285 schools reaching 11969 learners across the country. The study targets learners, educators and principals collecting responses on issues faced within the schooling system in addition to the mathematics and science assessment questions posed to grade 9 learners (Human Sciences Research Council n.d.). The survey of 2011 relates to four functionings valued by learners in terms of school dropouts, such as physical well-being, the infrastructure suitability of their learning environment, their ability to express their opinions freely and their ability to participate meaningfully in school. The particular questions linked to these functionings are described below in Table 74.

Functioning	Theme	Sub-Theme	Question
Being physically well	Malnutrition and Hunger	Number of learners affected by malnutrition	Students suffering from lack of basic nutrition
Being Taught in Suitable Infrastructure with necessary resources	Facilities and Resources	Available teaching aides	Teachers do not have adequate instructional materials and supplies
	Improve Physical Infrastructure	Ensure that the schools are of a suitable size	Classrooms are overcrowded
		School infrastructure maintenance	The school building needs significant repair
Free Expression of Opinions	Conducive school learning environment	Undisciplined learners	Uninterested students
	Comparative academic	Comprehension of classwork	Student achievement scores in mathematics and science

Functioning	Theme	Sub-Theme	Question
Meaningful participation in school	strength of the learner		Students' responses to each of the mathematics and science items administered in the study
		Understanding elementary level concepts	Students lacking prerequisite knowledge and skills
	Teacher Motivation	Lack of teacher motivation	I am content with my profession as a teacher
			I had more enthusiasm when I began teaching than I have now
			Teachers' job satisfaction
	Teacher Quality	Pedagogical skills, training and knowledge for teaching	By the end of this school year, how many years will you have been teaching altogether?
			What is the highest level of formal education you completed
		The ability to impart and instil knowledge, skills and values to the learners	Teachers have too many teaching hours

Table 74: Pertinent fields and data points from TIMSS 2011

The TIMSS 2011 survey scored highly achieving a score of 6.5 with a strong performance across dimensions except for applicability and currency. Similar to the analysis of other datasets, the data quality score follows closely from the assessment of the usefulness of the metadata. The reasons for the high performance of the TIMSS study compared to the SASAS survey is that the survey is based on an internationally agreed template with standards that are rigorously tested in numerous environments. Consequently, the definitions of terms and concepts are well detailed and supported with various pieces of documentation. Although the HSRC does not define data quality in the release documentation, various pieces of technical documentation are produced which detail how data quality is protected and ensured. In addition, the confidence interval of the data is described, the quality assurance processes are explained, the sampling framework is comprehensively documented, the questionnaire structure is discussed, the scope of the study is covered and the various concepts, definitions and statistical techniques applied are well discussed and referenced within the metadata provided to the user. Ultimately, the content, context and structure of the data files are highly detailed following the international TIMSS standards leading to a positive scoring across all elements and indicators of the useful metadata dimension.

The detailed metadata thereafter supports the positive scoring of dimensions such as data comprehensiveness, clarity, consistency and integrity. In this regard it was found that the details within the metadata match the values within the data. In terms of comprehensiveness, all the statistical units and data rules are documented, resulting in high data quality score.

Similarly, the data was found to be understandable to the data user as the terms used within the data matched the documentation. With data consistency, one finds that semantically and structurally the data provided corresponded to the metadata as all the indicators were positively assessing within the framework. Lastly, with respect to data integrity, it was found that the data quality constraints followed by the HSRC to ensure that the correct data were captured and the manner in which data quality breaches were handled, were adequately documented, thus reassuring the data user that the data made available is reliable.

In terms of data applicability, accuracy, accessibility and conciseness, there are some concerns that were identified. Data applicability was found to present the most concerns for data users as it could not be determined if there was sufficient public consultation regarding capturing the public's needs within the survey. Whilst such a practice is difficult to perform, no mention of any processes to test the local needs are included. Furthermore, no assessment is referred to within the documentation as to whether the public is satisfied with the production of the survey.

In terms of data accuracy, the limitation that was identified pertained to the testing the results of the survey against a comparative source where it was found that no such source is mentioned within the documentation. Whilst the dimension performs weakly in that single indicator, other indicators such as the adequacy of sampling techniques, imputation techniques, non-sampling statistical techniques and the adoption of data maintenance procedures were all found to be sufficiently detailed within the supporting metadata.

In terms of accessibility, the main limitations are linked to the application of the HSRC data embargo policy discussed in the review of the SASAS dataset. Ultimately, data is only released to the public two years after the initial release. This process results in the data losing relevance and blocks the public from conducting analysis on extremely pertinent trends. Considering the previous TIMSS was conducted in 2003, the infrequency of the data collection combined with the HSRC embargo severely diminishes the value of the data. Lastly, in terms of data conciseness, whilst the information is succinctly described, the scale of the survey sample restricts the data user from analysing the trends at a level lower than province in terms of geography.

In summary, the TIMSS strong performance signals that non statistical agents can produce data of high quality. Following the data quality requirements from the parental TIMSS body, the HSRC was able to produce data that scores highly in a series of dimensions. The data is extremely relevant, though is infrequently produced. Together with the restrictive HSRC data embargoes, the survey loses its impact due to its late release.

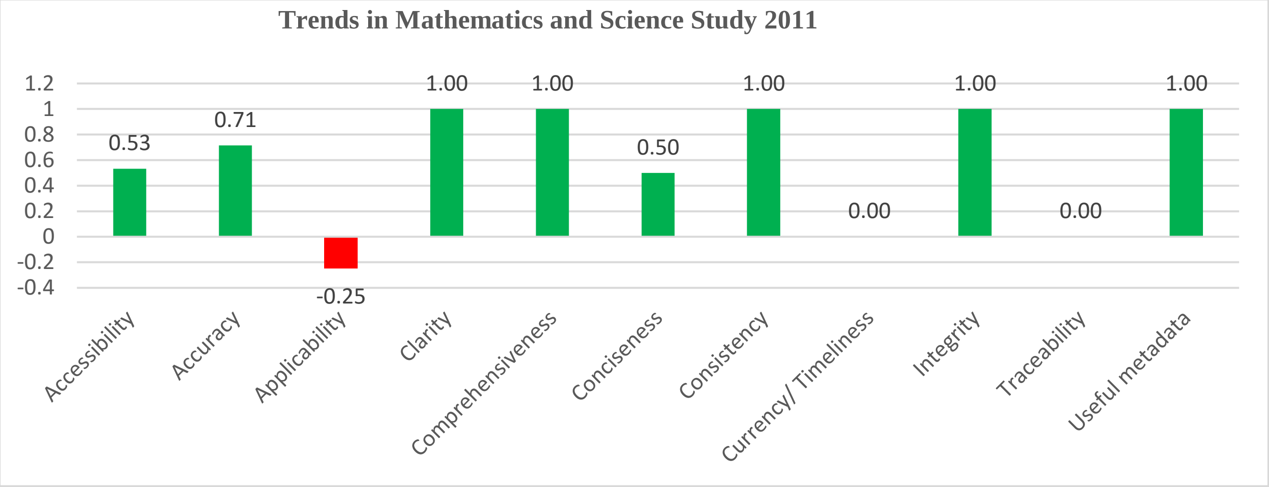


Figure 25: Data Quality Scores of the Trends in Mathematics and Science Study 2011

Youth Risk Behaviour Survey 2011

The Youth Risk Behaviour Survey (YRBS) was conducted last in 2011 by the South African Medical Research Council (MRC) together with the Human Sciences Research Council (HSRC). The survey updates the results of the previous surveys conducted in 2002 and 2008 in assessing the levels of risky behaviour by youth at a provincial level across the country. The survey's designers held the belief that risky behaviour amongst the youth is a determinant for the spreading of diseases such as HIV as the youth grow older (South African Medical Research Council n.d.). However, the study is also useful in assessing the various functionalities valued by learners as the survey is targeted to school going learners. The functionings that are related are physical well-being as well as the learner's meaningful participation in school. Table 75 identifies the relevant questions related to these functionings.

Functioning	**Theme**	**Sub-theme**	**Questions**
Being physically well	Teenage Pregnancy	Accessibility of contraception	Percentage of high school learners who used various methods of contraception by gender, race, grade, age and province
			Percentage of high school learners who always use condoms, who had either been pregnant or made someone pregnant, and who has a child/
	Malnutrition and Hunger	Number of learners affected by malnutrition	Percentage of high school learners who are undernourished and over-nourished by gender, race, grade, age and province
Meaningful participation in school	Household and	Crime, violence, substance abuse and other	Percentage of high school learners who carried a weapon by gender, race, grade, age and province

	Community Environment	negative environmental factors	Percentage of high school learners who engaged in violence related behaviours by gender, race, grade, age and province
			Percentage of high school learners who perpetrated or suffered partner violence and coerced sex by gender, race, grade, age and province
			Percentage of high school learners who engaged in violence related behaviours on school property by gender, race, grade, age and province
			Percentage of high school learners who engaged in, watched or tried to stop a physical fight on school property by gender, race, grade, age and province
			Percentage of high school learners who were driven by a driver who had been drinking alcohol, who drove after drinking alcohol, who walked alongside
			Percentage of high school learners who use alcohol by gender, race, grade, age and province
			Percentage of high school learners who used cannabis (dagga) by gender, race, grade, age and province
			Percentage of high school learners who use other drugs by gender, race, grade, age and province

Table 75: Pertinent fields and data points from the YRBS 2011

The YRBS of 2011 scores highly in some data quality dimensions. The positive, but weak performance in the usefulness of the provided metadata underlines the comparatively mediocre data quality score of 0.97 that was attained. The weakness of the metadata is attributable to a lack of detail describing the structure of the data files produced where no data file codebook is released to the public, data quality breaches are not discussed, issues of scope are excluded, the metadata lacks references to guiding standards used in setting up the survey and the particular hardware and software requirements for analysing the survey are left out from the documentation. These concerns diminish the various positives found in the metadata that relate to details about the sampling framework, the incorporation of a pilot study to test the results and the various validation processes that were applied and documented.

These weaknesses in the metadata translated into weaknesses in accessibility, applicability and data currency. In terms of data accessibility, the survey was found very weak due to two essential reasons. Firstly, the metadata does not describe the details of the fields and tables released and secondly, neither the MRC nor the HSRC makes the data accessible on their websites and the organisations do they offer any details regarding how to go about accessing the public information. Data currency is assessed negatively for a similar reason considering the MRC and HSRC have not actually released the data files to the public. Due to the lack of data made available, currency can only be negatively assessed. Data comprehensiveness is assessed weakly for similar reasons. As no codebooks were provided the details about the data files and the detail of the statistical units related to each field cannot be determined to be complete, hence the weak data quality score for the dimension. With respect to data applicability, the study suffered from a lack of documentation pertaining to the public's perception of the data release as well as the inclusion of extra weights included within the dataset that were not described within the dataset.

Despite, the lack of published data files, the other dimensions of data quality (clarity, conciseness, consistency, accuracy and integrity) are assessed positively. Data clarity was rated strongly as the terms and concepts were clearly defined and used in the datasets. In terms of data conciseness, all the fields in the questionnaire are found to be relevant for the data users. With respect to consistency, the questionnaire remains consistent with the 2002 and 2008 surveys with a few additional questions. Furthermore, the reported data matches the terms used in the questionnaire in the provided data files. Data accuracy was assessed strongly due to the inclusion of the data sampling calculations, the reported confidence interval and the incorporation of a pilot study to test the results of the survey. The weakness that was identified pertained to the external audit of the MRC. Whilst the organisation is audited, the audit reports do not mention the YRBS specifically. Lastly integrity was assessed positively as the metadata included references to input constraints that were applied within the electronic questionnaires, however no mention was included in the metadata regarding how data quality breaches were handled, if found post data capture.

The major concerns in data quality related to the YRBS is the lack of a clear data dissemination policy. Whilst the HSRC assisted the MRC in data capture, neither organisation has taken responsibility to release the data to the public. In this regard, general data users would fail to assess the survey at all. However, after contacting the principle investigators responsible for the survey, the dataset was assessed using the PDQAF. From this assessment, one finds that the lack of a provided codebook, and the limitations of a single data format restricts the usability of the dataset. Care must be taken to link fields within the SPSS data files to specific questions on the survey. However, in review of the documentation that is provided, it is noted that various quality controls are put in place resulting in a strong performance in terms of accuracy, clarity, consistency and conciseness. The individual scores per data quality dimension are described in Figure 26

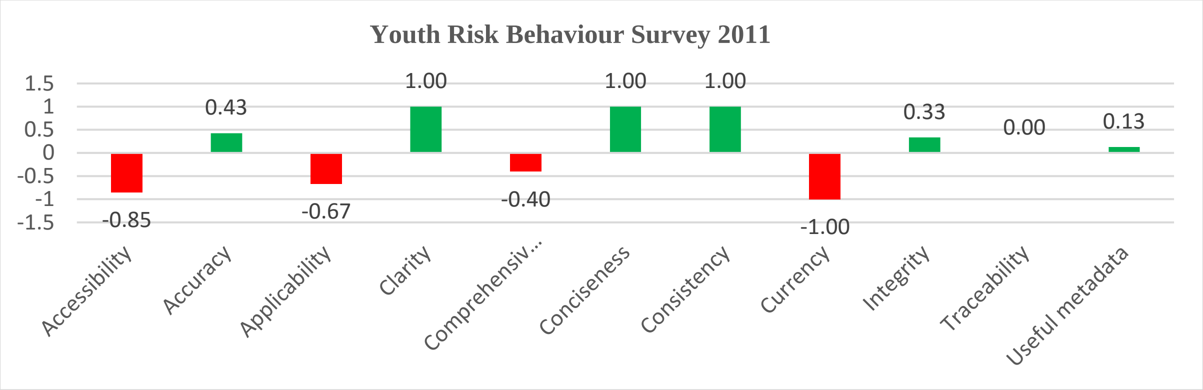


Figure 26: Data Quality Scores of the YRBS 2011

South Africa Demographic and Health Survey 2003

The South African Demographic and Health Survey (DHS) of 2003 was released in 2007 by the South African Medical Research Council (MRC) together with the South African National Department of Health. The survey examines the changes in health status of the population since the advent of the previous survey which was conducted in 1998. The factors that are explored within the survey included vulnerabilities of the population, access to health services, and status related information such as pregnancy, child health, fertility and knowledge of contraception. These factors are very useful to assess physical well-being as well as the learner's financial security, their learning environment and the learner's participation in school. The specific questions that were selected as relevant within this assessment are highlighted in Table 76.

Functioning	**Theme**	**Sub-theme**	**Questions**
Being financially secure	State funded programmes	Provision of state grants targeted at poor households	Grants and pensions, recent injuries
Being in a conducive home learning environment	Negative attitude to learning in the household, community	Education attainment of the head of household and associated community	Educational level of female and male household population
			Level of education
Being physically well	Malnutrition and Hunger	Number of learners affected by malnutrition	Nutritional status of children
			Nutritional status of children at birth
		Nutrition intake of learners	Micronutrient intake among children
			Age at first marriage

Functioning	Theme	Sub-theme	Questions
	Menstruation and Female Maturation	Gender biased communities	Median age at first marriage
			Polygyny
	Teenage Pregnancy	Accessibility of contraception	Condom use at first sex among young women and men
			Current use of contraception
			Current use of contraception by background characteristics
			Current use of contraception by women's status
			Ever use of contraception
			Higher risk sex and condom use at last higher risk sex in the last year among
			Multiple sex partnerships among young women and men
			Number of children at first use of contraception
			Sexual activity and condom use in last 12 months
			Source of contraception
			Timing of sterilization
		Pregnant <20 years	Knowledge of contraceptive methods
			Teenage pregnancy and motherhood
Meaningful participation in school	Comparative academic strength of the learner	Understanding elementary level concepts	Literacy

Table 76: Pertinent fields and data points from the DHS 2003

Although the DHS of 2003 follows the international guidelines of the Demographic and Health Survey Programme funded by the United States Agency for International Development (USAID) the metadata makes little mention of the connection to the international body in terms of the various standards and principles that are followed when carrying out the survey. Furthermore, the survey results are released by the MRC, however the released documentation does not provide any details about the structure of the datasets and neither is the dataset actually released. All data is made available within the findings report of the survey. These limitations are exhibited in the poor data quality score of the survey (-4.90) as a whole and the low score regarding the usefulness of the metadata.

The difficulties in assessing the metadata result from the packaging of the metadata within the findings report as an appendix. This results in a lack of detail provided in the report regarding the file layout, how data quality faults are addressed and resulting from the exclusion of definitions pertaining to fields that are used within the findings report. Within the metadata section of the findings report, there is a discussion of how field-workers are trained to conduct the survey correctly, the sampling framework is discussed as well within the inclusion of the response rates confidence intervals and sampling errors. These details signal that data quality is considered, just not reported on effectively.

The lack of published data and codebooks, highly limits the assessment of data comprehensiveness, accessibility, clarity, conciseness and consistency. Data comprehensiveness cannot be assessed as the lack of data does not permit one to explore whether all required statistical units or data rules were collected. Furthermore, the data rules can only be assessed in terms of the instruction included within the questionnaire which is also provided as an appendix to the data findings report. The data accessibility dimension is scored negatively precisely because the detailed data cannot be accessed due to the lack of available data and the lack of provided instructions as to how one can request the DHS information. Data clarity is also limited because concepts and fields are not defined within the metadata, whilst conciseness cannot be assessed as the data files are unavailable for review. Lastly data consistency is rated negatively as the dataset is not available and the terminology within the dataset cannot be compared to the previous iteration of the survey conducted in 1998 (which is also not published).

Other limitations to data quality aside from the lack of available data to test relates to data integrity and currency. Data currency was negatively scored as the MRC has not published the schedule pertaining to the release of the information. Furthermore, the survey results were only released 4 years after collection, compounding the poor data currency assessment. Data integrity is scored poorly as the documentation that was released does not discuss the input constraints applied to protect the data collected and neither is the process for addressing such quality breaches included in the metadata. Lastly, the data aggregations applied within the findings report are not discussed, further weakening the score of the data integrity dimension.

On the positive side, the assessment of data accuracy scores strongly due to the reported information of sampling techniques, error rates and confidence intervals pertaining to tests performed on the sourced data. Each of the factors are drawn from the discussion within the metadata but cannot be assessed physically due to the lack of data for the user to work with directly. The only other weakness highlighted pertains to the lack of a corresponding data source that can be used to verify the results the of the survey.

In review of the survey in its entirety, the lack of available data files for the public to review and work with is a major concern. The discussion within the metadata pertaining to the sampling frame and other statistical calculations that were conducted alludes to data quality practices that were prioritised internally within the MRC, however these practices are not fully documented. The inclusion of metadata as an appendix to the findings report suggests that metadata is not viewed as a priority within the MRC. The data release would benefit from a report structured purely along the lines of the needs of metadata which pays attention to the organisation's various requirements for data quality. Without such documentation, it is assumed by the data user that data quality is not a priority for the MRC. The individual data quality scores per dimension are presented in Figure 27.

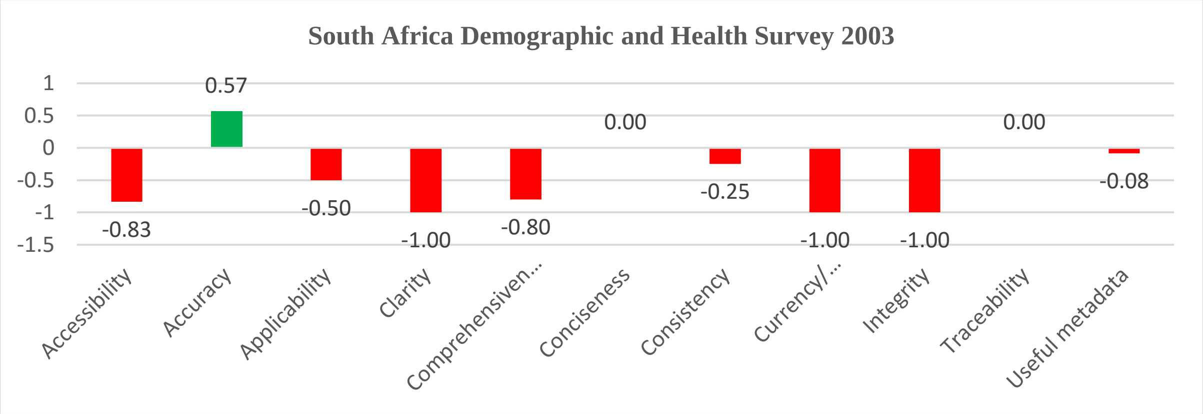


Figure 27: Data Quality Scores of the SA Demographic and Health Survey 2003

Crime Statistics

The South African Police Services (SAPS) release on their website, statistics on the incidents of crime occurring across the country. No supporting metadata is published explaining to the data user of the details of the data (South African Police Services n.d.). The information is valuable as it describes the reality pertaining to household and community environment that learners face in terms of crime, however none of the data quality dimensions can be assessed in detail due to lack of information. This results in the poorest data quality score attained of -7.29.

Functioning	Theme	Sub Theme	Specific Crime Statistics Data
Meaningful participation in school	Household and Community Environment	Crime, violence, substance abuse and other negative environmental factors	Contact crimes (crimes against the person)
			Contact-related crimes
			Property-related crimes
			Crime detected as a result of police action
			Other serious crimes
			Subcategories of aggravated robbery

Table 77: Pertinent fields and data points from the South African Crime Statistics

In review of the data quality assessment, the weak score attained is primarily due to the lack of any documentation to support any reasoning that data quality has been prioritised by the organisation. The lack documentation means that no assessment of data rules, sampling, quality checks or statistical calculations can be performed. The trustworthiness of the data is dependent on the user's perceptions of the organisation alone. As the police services are not primarily data collectors, one cannot assume that data quality processes are applied. Whilst the data provided is useful to the data users, not much effort is made to report on the data quality practices. Greater documentation needs to be made available to help the data user make an assessment of the associated level of data quality to help build their level of trust when referring to such data. The specific scores per dimension are provided below in Figure 28.

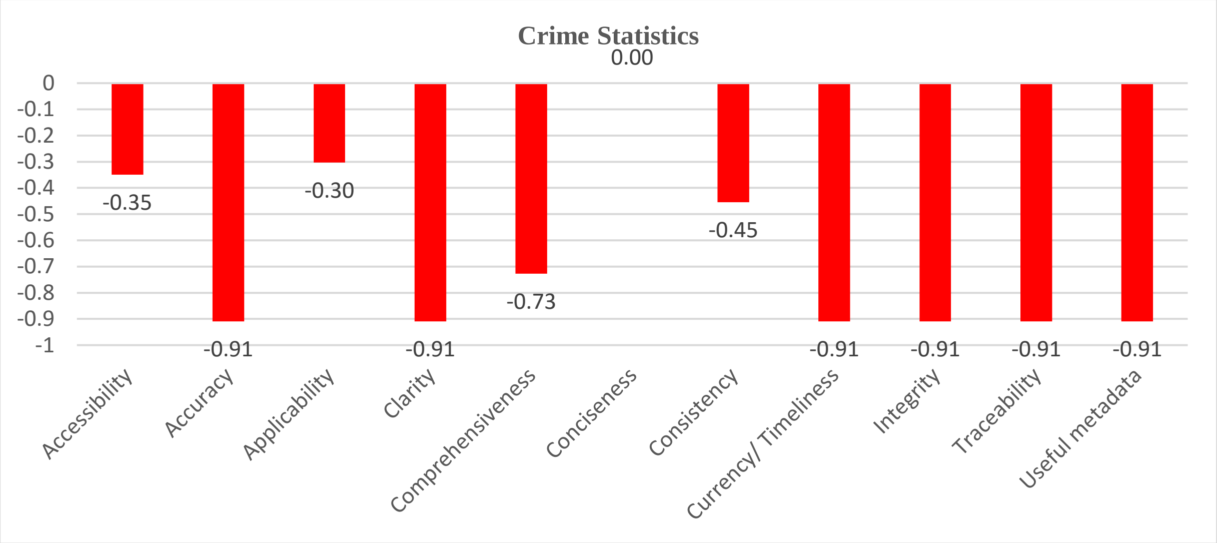


Figure 28: Data Quality Scores of the SA Crime Statistics

Appendix 3
Data Quality Assessment of Brazilian Datasets

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
Is the Metadata useful to the data user?	Does the metadata convey the contents of the dataset?	Are concepts and definitions and classification provided within the metadata to describe the underlying data?	Are definitions made available within the metadata?	A dictionary is provided for each questionnaire. Fundamental concepts are discussed in detail	Concepts and definitions are discussed in detail in the Census results documentation	A Data Dictionary is provided on the PNDS Website	A Data Dictionary is provided with the data release	No metadata is provided
			Are deviations from standards reported?	Not applicable -Standard deviations do not apply to the dataset. Only the geography codes need to be consistent with country standards. The codes are discussed within the geography	All concepts are well defined throughout the various pieces of documentation	Concepts from the International Demographic and Health Survey Programme and internationally accepted	International Standards used in the data collection are not referenced within the metadata	No metadata is provided
		Is the metadata up to date?	Is a document register regularly maintained in line with the data collection strategy?	The documentation is provided for each survey for each annual iteration of the survey	The documentation is provided for each census	The documentation is provided for the particular PNDS survey alone. Frequent changes are not applicable	The provided metadata is refreshed for the annual data collection	No metadata is provided
		Are all tables and fields defined?	Is the purpose and definition of each table and field within the dataset reported on?	The fields of each data table are defined with descriptions for each option	Data is provided in various table structures on the IBGE Website at different levels of aggregation, not every restructure of census data is accompanied by documentation with definitions of the terms that are used.	Fields are described within the data extract. Individual table structures are not discussed	Fields and tables are described within the metadata	No metadata is provided
				Differences between similar fields are described	Fields used in the census are consistent with their use in Household and other demographic surveys of Brazil	Fields used are consistent with international naming conventions, although are translated into Portuguese	Fields used are largely consistent with international norms but the norms are not directly referenced	No metadata is provided
		Is the geographic distribution clearly defined?	Is the geographic level of data granularity described within the metadata?	The granularity is documented; information is provided to school code level	The granularity is documented; information is provided to sub-district level via the IBGE website	The Geographic granularity of the data output is not detailed within the documentation	The Geographic granularity is described within the data dictionary	No metadata is provided
				Not applicable - Granularity changes are not applicable	Geographic boundaries are well documented	Geographic boundaries are well documented	Geographic structures and codes are referenced within the data dictionary	No metadata is provided

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
	Does the metadata describe the context of the dataset?	Does the metadata describe the data quality practices applied when producing the dataset?	Does the organisation define data quality within the metadata?	The metadata includes reference tests matrices to guide the sampling framework, using literacy and numeracy results as benchmarks for data collection.	The methodology documentation discusses the various validation techniques that were applied	The methodology documentation discusses the sampling plan, sample structure, the results of a Pilot Study and the weighting calculations used	No data quality discussion is included within the metadata	No metadata is provided
			How does the organisation measure data quality within the metadata?	The metadata is not explicit about how data quality is measured, if at all	The metadata is not explicit about how data quality is measured, though it does describe the various techniques to promote quality	The metadata is not explicit about how data quality is measured	No data quality discussion is included within the metadata	No metadata is provided
			How does the organisation process data concerns? Is this discussed in the metadata?	Uses Item response theory and correction tests	Methodology discusses the validation and integration checks that are put in place.	Scenarios are provided within the methodology regarding techniques to use during an interview to ensure data quality	No data quality discussion is included within the metadata	No metadata is provided
			Are data quality controls discussed within the metadata?	The correction tests applied to the data are discussed within the metadata	A process flow is discussed within the methodology discussing how data quality concerns were addressed.	Methodology discusses the validation and constraints applied during data capture	No data quality discussion is included within the metadata	No metadata is provided
		Does the metadata comprehensively describe the data collection methodology?	Are the statistical techniques employed suitably documented?	The sampling framework is not discussed in detail but the correction tests are discussed	Effort is made to target all households within the country and this is discussed within the methodology	The theoretical development of the weight calculation is provided within the methodology	The data collection methodology is not discussed	No metadata is provided
			Are all data sources used documented?	Not applicable	Not applicable	Not applicable	The data collection methodology is not discussed	No metadata is provided
			Is the sampling selection framework and decisions taken adequately documented?	The sampling framework is not discussed in great detail	A discussion regarding the enumeration and processing of the small area layer is provided within the methodology	The sampling plan and the expansion of the sample are detailed within the methodology	The data collection methodology is not discussed	No metadata is provided
			Are the techniques for ensuring anonymity provided?	The target audience is discussed but not individually referenced. Data is aggregated to a school level and not a learner level	Data is aggregated to a Sub-District level, as well as other higher levels of geography	Data is aggregated over individual households when released	Learner names are not provided but unique learner IDs are provided	No metadata is provided
			Is the questionnaire design documented?	The structure is well documented, as well as the SPSS data structure	The structure is well documented	The questionnaire structure is published	The questionnaire structure is not provided	No metadata is provided
			Are the norms and standards regarding the data collection discussed within the metadata?	The norms are not discussed	Comparison to international norms (connected to Statistical Institute of Mercosul) are discussed within the Planning Chapter of the Methodology.	Comparison to international norms (Demographic and Health Survey Programme) are discussed within the Methodology.	The data collection methodology is not discussed	No metadata is provided

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
			Is the scope of the data collection documented?	The target audience of the survey is specified	The scope is outlined in the methodology	The in an out of scope sections is outlined in the methodology	The scope of the survey is not discussed within the metadata but the INEP website notes that all schools are targeted. The range of questions covered in the survey are not discussed	No metadata is provided
		Is there an accompanying findings report?	Is a data output report provided together with the dataset publication?	A detailed findings report is provided	A detailed findings report is provided	A detailed findings report is provided	A detailed statistical synopsis report is provided; however, no discussion of the results is provided	No findings report is provided apart from the statistics
		Is the metadata clear and understandable to the data user?	Does the metadata concisely and comprehensively describe how the dataset is produced?	The core terms are defined within the documentation	Concepts and definitions are discussed in detail in the Census results documentation in simple language	Concepts and definitions used are discussed in the results documentation and are understandable	Concepts and definitions used are understandable, but no analysis of statistical tables are offered to add insight to the reported numbers	No metadata is provided
				The sequence of information presented in the documentation is provided well	The sequence of information presented in the documentation is provided well	The documentation excludes details about the structure of the data structures	The documentation is presented to support the user's analysis of the data	No metadata is provided
				The statistical techniques are described and referenced well	The statistical techniques are described and referenced well	Each section in the methodology carries references	The documents do not refer to external documentation apart from the provided statistics	No metadata is provided
				Documentation is directly applicable to the data	Documentation is directly applicable to the data	Documentation is directly applicable to the data	Documentation is directly applicable to the data	No metadata is provided
	Does the metadata explain the structure of the dataset?	Does the metadata document the structure of the dataset?	Is the physical layout (data tables, fields, database) of the data structures documented?	The physical layout is described	Various structures of Census data are provided on the IBGE website without documentation of the physical layout	The table structures are not detailed	The data structures are well documented	No metadata is provided
			Are the hardware and software requirements detailed regarding use of the data?	The software requirements are discussed. The hardware requirements are ommitted but are not essential	The software and hardware requirements are discussed.	Software and hardware requirements are not discussed	Software requirements are discussed	No metadata is provided
			Does the metadata explain how to navigate and use online databases if relevant to the data publication	The data is not provided via an online data portal	An interactive database tools is not utilised to extract the data	An interactive database tools is not utilised to extract the data	An interactive database tools is not utilised to extract the data	No metadata is provided
Is the data comprehen sive for the data user’s	Does the dataset contain all required statistical units?	Are all expected statistical units populated?	Is each combination of statistical units represented within the data as per the required scope of the dataset?	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
requirements?		Does the dataset structure match the metadata outline?	Does the dataset layout match the metadata in terms of the table, field names provided?	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided
		Does the dataset reasonably convey the definitions, scope, classification, valuation and timeline as specified within the metadata?	Do values provided within a field broadly match the definition?	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	No definitions are provided
			Are the provided Statistical units matching the scope of the dataset?	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided
			Do the categories within the data match the specified classification?	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided
			Does the data match the data type of the particular fields?	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided
			Does the time periods within the data match the expected time period?	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided
	Are all data rules met within the dataset?	Are there blank data entries where the data rule mandates an entry?	As per the data rules, are all necessary fields provided and populated?	All necessary values are provided	All necessary values are provided	All necessary values are provided	All necessary values are provided	The data extract is populated; however it can be determined what the rules may be
		Are all mandatory attributes populated?	Are Mandatory fields populated?	All mandatory fields are populated	All mandatory fields are populated	All mandatory fields are populated	All mandatory fields are populated	There are no blank entries in the statistical reports
		Are all inapplicable attributes left blank?	Are inapplicable attributes blank where appropriate	Inapplicable fields are empty	Inapplicable fields are empty	Inapplicable fields are empty	Inapplicable fields are empty	unable to determine which are inapplicable fields due to the lack of metadata
Is the data accurate for the data user’s purpose?	Are the data coverage methods adequate?	Has the dataset adequately applied data sampling techniques?	Are the standard error, the coefficient of variation, the confidence interval and the mean square error calculation in line with international norms and standards	The standard error, variation coefficients, confidence intervals and the mean square error are not provided	The data sampling calculations are internationally accepted	The data sampling calculations are internationally accepted	Data sampling calculations or standard errors, etc are not provided in the metadata	No metadata is provided

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
		Are imputation techniques adequately applied	Is the response rate too low and is imputation rate too high?	The imputation rate is not discussed	Improved techniques were used to reduce the need for imputation techniques by adopting greater electronic questionnaire tools and detailed logic tables to guide imputation. Imputation has been found to be low. A 95% Confidence level was achieved in line with international norms	A 95% confidence interval was achieved across the survey in line with international norms	Imputation techniques are not discussed within the metadata	No metadata is provided
		Has the dataset adequately applied non sampling techniques	Is the frame coverage, duplication in the frame, number of statistical units out of scope, misclassification errors, measurement errors, processing errors and model assumption errors data calculations in line with international norms and standards?	Reference test matrices are applied and their results are provided and the differences are within accepted ranges	Frame coverage is suitable	Frame coverage is suitable	Non sampling techniques are not discussed	No metadata is provided
	Is the data suitable for reporting?	Does the provided data correspond with a comparative source?	The aggregated data of the dataset equates to the comparative data source	The reference test results were comparative	The census result compared adequately with the Census Pilot Survey	The survey compares suitably to the pilot study	No comparative source or pilot study is identified.	No comparative source or pilot study is identified.
		Has a downstream source been quality assessed?	Was the primary data source assessed in line with the Public Data Quality Assessment Framework	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
		Are the adopted maintenance procedures suitable?	Has the sample frame been regularly updated and managed?	According to the metadata, the sample frame is assessed at the start of the study	According to the methodology, the sample frame is evaluated at the start of the census	According to the methodology, the sample frame was established on initiation of the project	No references to a sample frame are provided	No references to a sample frame are provided
			Does the organisation conduct regularly quality assurance procedures?	Each survey connected by INEP follow similar practices	Each survey connected by IBGE implement data quality practices	The survey is not regularly collected	Quality assurance are weakly referred to on the INEP website, where schools are invited to register on the site and provide feedback on data collections. The process is not documented clearly	No metadata is provided to make a determination
			Does the organisation conduct regular data audits and are the errors identified acceptable?	INEP is regularly audited and provides Management Reports	The IBGE is audited externally	The Health Ministry is externally audited	INEP is regularly audited and provides Management Reports	SINESP is regularly audited

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
Is the data clearly understand able to the data user?	Is there consistency between terms used and the naming convention?	Is there consistency between the terms used and the organisation's definitions?	Are concepts, definitions, classifications and standards applicable to the dataset made available?	Concepts are consistently applied across INEP	Concepts are consistently applied across IBGE	Concepts are consistently applied	Terms are consistently used across the different tables and datasets	There is no metadata to determine if the definitions are employed correctly
			Are the terms used based on agreed company definitions?	Terms follow international norms on literacy, although documentation requires English translation	Terms follow Mercorsul Norms	Terms follow Demographic and Health Survey Programme Prescripts	Definitions are clearly provided by INEP and accompany the dataset publication	There is no metadata to determine if the definitions are employed correctly
Is the data applicable to the data user?	Does the dataset meet the needs of the data users?	Does the dataset meet the requirements of the data users?	Have the public's data requirements been determined and have they been made available?	It is difficult to assess the public' data requirements for a literacy assessment	Cannot be determined if all requirements of the public are captured within the Census	Cannot be determined if all requirements of the public are captured within the survey	The public's data requirements have not been determined or are not made available	The public's data requirements have not been determined or are not made available
			Do the identified data requirements correspond with the purpose of the dataset?	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable
			Is the public satisfied with the contents of the dataset?	School directors assess the data and can appeal where there are differences in released data	Imputation was reduced and data quality satisfaction increased amongst data verification agents	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented
	Does the dataset contain beneficial information for the data user?	Does the dataset contain unnecessary/superfluous details not required by the data user?	All sub-components of the dataset must be examined to determine the granular subunits are relevant to the data user	Technical information and unimportant data has been excluded	Technical information and unimportant data has been excluded	Technical information and unimportant data has been excluded	Technical information and unimportant data has been excluded	Reported data is pertinent to crime reporting
Is the data concise enough for the data user?	Is the data concisely presented to the public?	Is the data to the point?	Does the data contain unnecessary elements for the data user?	All fields provided are relevant for literacy assessment	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant
			Does the data provide a sufficient level of granularity for the data user?	School level detail is very relevant for research purposes	Information lower than Sub-District may be required amongst some communities. This is currently not freely available.	Individual responses are provided	Individual responses are provided	Data users may require greater granularity per thematic area or geographical location
Is the data consistently represented to the data user?	Is the data semantically consistent?	Are data rules and definitions applied consistently to the dataset?	Does the data provider follow a common methodology (international standard) when producing the dataset?	The questionnaire construction is consistent with international norms	The questionnaire construction is consistent with Mercorsul norms	The questionnaire construction is consistent with Demographic and Health Survey Programme	The data collection methodology is not discussed	The data collection methodology is not discussed
			Does the metadata describe changes to the methodology and/or rules, definitions that are used?	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms	The data collection methodology is not discussed	The data collection methodology is not discussed
	Is the data structurally consistent?	Are data values consistently	Are data values over time are reported consistently?	The data matches the results documentation as per the audit reports	The data matches the results documentation	The data matches the results documentation	The data matches the results documentation	Data is consistently reported over the period of available data

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
		reported to the public?	Does the data provider reasons for changes in data values?	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms	If there were changes, it is not documented and therefore one is unaware if there were issues
Is the data current enough and released in a timely manner for the data user?	Is the dataset punctually released?	Is the dataset released in a punctual manner?	Is the average time between the end of the data collection and the data release within recommended time-frames?	The data was release two years following the legislated ordinance for the data collection	Data was collected between August and October 2010 and was fully released by November 2011	The time taken to publish the survey after completion of data collection activities is not published	The data is released annually	Crime statistics are released two years after collation annually
			Does the organisation release a report on the time frame for data releases?	Preliminary 2015 results have been released, results are getting vetted by school directors before the next official release. Not officially stated when the release will occur	The census is carried out every 10 ten years	The survey is carried out every 10 ten years	A time frame for reporting is not provided	A time frame for reporting is not provided
Is the data accessible to the data user?	Are data access controls too stringent?	Are the applicable data users granted an appropriate level of data access?	Are the access rights clearly and uniformly applied ensuring that the correct individuals are granted access?	Freely available - no limitations. Respondent information is aggregated over	Freely available - no limitations. Hard copies of meta data are charged for	Freely available - no limitations.	Freely available - no limitations.	Freely available - no limitations, detailed information is unavailable
			Can users easily find directions on how to access the required information?	Not applicable	Information is available on the IBGE website.	Information is available on the Health Ministry Website	Information is available on the INEP Website; the statistical synopsis is not clearly indicated where it is provided	No supporting documentation is provided
			Does the data provider respond to data requests in line with international standards on responsiveness	Not applicable	Information is available on the IBGE website.	Information is available on the Health Ministry Website	Data is available for download	Data is available for download; no additional user support is provided for more detailed information
	Is a suitable granularity of data accessible to the public?	Are access controls applied to the correct fields and categories?	Are access controls applied to the correct selection of fields and or categories?	data is available for download	data is available for download	data is available for download	Only learner names are protected on the published dataset	Granular data is not made available, there are no mechanisms to request greater detail.
			Adequate levels of anonymity need to be applied whilst ensuring that the user's needs for data are considered	Data has been anonymised	Data has been anonymised	Names of respondents are removed	Names of learners are removed	Data is aggregated to municipal level ensuring anonymity
			Does the metadata describe the levels of detail that are accessible?	Learner level information is not provided, only available internally within INEP	The results documentation discusses the detail available within the Census, supporting codes and descriptions are available for download	The metadata does not describe the geographic details or the file structures	The metadata describes the data structure and detail well	No metadata is provided

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
	Is the data and metadata provided in a meaningful and useful manner?	Is data and metadata provided in a useful and meaningful manner?	Is the data dissemination strategy shared with the public?	The website highlights where data can be downloaded	Long form and short forms of the metadata are available online to assist the public user to quickly identify data quality concerns	There is not published data dissemination strategy	There is no published data dissemination strategy, but the data is released publicly	There is no published data dissemination strategy, but the data is released publicly
			Is the data release scheduled?	The next release is not dated	The next release is not dated	The next release is not dated	The next release is not dated	The next release is not dated
			Does the metadata provide guidance on how to access and retrieve data using the available mediums?	Data access is discussed within the metadata	Tables are provided without guidance on how to use the various mediums available	Tables are provided without guidance on how to use the various mediums available	The metadata provides steps on how to access and retrieve data	No metadata is provided
			Is the data dissemination strategy frequently updated?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
			Does the data dissemination strategy take heed of the needs of the data users?	Data is easily available	Database tools could be adopted to improve usability when interacting the data. Data is primarily made available by static files	Database tools could be adopted to improve usability when interacting the data. Data is primarily made available by static files	Data is published with no dissemination strategy documented	There is no data dissemination strategy that is published
	Is the channel of data delivery appropriate for the data user?	Is the channel of data delivery suitable for the needs of the user?	Are hardware and software requirements clearly communicated to the data user?	Software requirements are discussed	Software requirements are discussed	Software and hardware requirements are not discussed	Software requirements are not explicitly discussed	Software and hardware requirements are not explicitly discussed
			Have the user's preferred means of accessing data been taken into consideration when developing the data access channel?	Not practical	The channel could be improved	The channel could be improved	The channel could be improved with database tools, but Excel is a common preference for distributing data	The channel could be improved with database tools, but Excel is a common preference for distributing data
			Are the disseminated datasets accessible in multiple data formats?	Provided in multiple formats	Provided in multiple formats	Provided only in SPSS	Provided only in Excel	Provided only in Excel
			Where database tools are provided, is the mode of access understandable and simple to operate?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
			Are user manuals provided to aid data access to database tools?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
			Where external staff resources need to be contacted to request dataset, are such staff reachable?	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
Does the data have integrity when	Are data quality constraints suitable for the	Do the data quality constraints applied within	Do the data quality constraints follow international best practices?	The data excludes non-applicable values	Great attention is directed to data quality constraints introduced to the data collection strategy	Data input constraints are applied to the electronic form capture used in the survey	Data quality constraints are not discussed	Data quality constraints are not discussed

Dimension	Element	Indicator	Focal Issue	ANA 2014	IBGE Census 2010	PNDS 2006	School Census 2015	Crime Statistics
consumed by the data user?	needs of the data user?	the data system sufficient to manage data integrity?	Are the data rules suitably captured within the data quality constraints?	Only data values as discussed within the metadata are provided	Only data values as discussed within the metadata are provided	Only data values as discussed within the metadata are provided	Data quality constraints are not discussed	Data quality constraints are not discussed
			Is the process to address data quality breaches discussed within the metadata?	The process to query a result when the data is preliminarily release is not discussed within the metadata but only on the INEP website	The process for resolving data quality breaches are detailed within the Methodology	The process for resolving data quality breaches are not detailed within the Methodology	Data quality constraints are not discussed	Data quality constraints are not discussed
	Does the data provider introduce the most appropriate statistical and methodological approaches?	Are the data transformation techniques suitable?	Are the applied data transformation techniques applied within the bounds of international best practices?	The aggregations applied to the data are adequate meeting the reference tests applied	The data transformations follow Mercosul norms adopted by the IBGE	The data provided is not aggregated	The data provided is not aggregated	There is no metadata to assess the data transformation techniques applied
			Are the data transformation methodologies detailed within the supporting metadata?	Aggregation matches the metadata	Aggregation matches the metadata	The data provided is not aggregated	The data provided is not aggregated	There is no metadata to assess the data transformation techniques applied
Is the data traceable?	Does the data provider track the data source?	Does the organisation track data transformations ?	Does the organisation document the data sources and transformations which are used within the metadata?	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	There is no metadata to assess the quality of the source systems that are used
			Has supporting information been vetted using the public data quality assessment framework?	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	There is no metadata to assess the quality of the source systems that are used

Appendix 4

Data Quality Assessment of Indian Datasets

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
Is the Metadata useful to the data user?	Does the metadata convey the contents of the dataset?	Are concepts and definitions and classification provided within the metadata to describe the underlying data?	Are definitions made available within the metadata?	Concepts and definitions are discussed in detail in the Census Handbook	Concepts and terms are defined within an Educational Planning Guidebook	No metadata is provided - terms are alluded to in the categorisation and definitions of crimes are provided, but no specific references to fields and tables are provided	A Data Dictionary is provided on the DHS Website
			Are deviations from standards reported?	Many terms used within the Census are specific to the Indian context only, although, these terms are described within the metadata	Effort is being made to align the data collection to international best practices, but the deviations from best practices are not referred. Only copies of signed agreements are published on the DISE website.	No discussion on standards are provided	Concepts from the International Demographic and Health Survey Programme and internationally accepted
		Is the metadata up to date?	Is a document register regularly maintained in line with the data collection strategy?	The documentation is provided for each census	The guidebook is outdated. Was produced in 2003	The documentation is from 2005	The documentation is provided for each annual survey. Frequent changes are not applicable
		Are all tables and fields defined?	Is the purpose and definition of each table and field within the dataset reported on?	Data is provided in various table structures on the IBGE Website at different levels of aggregation. Supporting documentation of each table structure is not provided.	The metadata discusses all the concepts in general. Individual surveys, tables and fields are not described	No metadata is provided	Fields are described within the data extract. Individual table structures are not discussed
				Fields used in the census are consistent with their use in Household and other demographic surveys of India	Fields in tables follow from descriptions in the guide book	No metadata is provided	Fields used are consistent with international naming conventions, although are translated into Portuguese
		Is the geographic distribution clearly defined?	Is the geographic level of data granularity described within the metadata?	The Geographic levels of the data clearly described	The Geographic granularity is not described within the handbook	No metadata is provided	The Geographic granularity of the data output is not detailed within the documentation due to the implementation of the survey internationally
				Changes to boundaries are reported on the website	The Geographic granularity is not described within the handbook	No metadata is provided	Geographic boundaries are not well documented
	Does the metadata describe the context of the dataset?	Does the metadata describe the data quality practices applied when	Does the organisation define data quality within the metadata?	The documentation does not define data quality but refers to some data quality processes that were put in place such as setting up a pilot study to determine the most appropriate set of questions for the census	A 5% data sampling technique is adopted to check the data that is collected. All states must conduct random checking to ensure data accuracy.	No metadata is provided	The methodology documentation discusses the sampling plan, sample structure, the results of a Pilot Study and the weighting calculations used

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
		producing the dataset?	How does the organisation measure data quality within the metadata?	The metadata is not explicit about how data quality is measured.	Data quality across all schools is not measured in the full datasets released per year	No metadata is provided	The metadata is not explicit about how data quality is measured
			How does the organisation process data concerns? Is this discussed in the metadata?	The methodology does not provide details about how data concerns are addressed	Random sample reports of specific district's data checks are provided on the DISE website to caution districts against incorrect data collections.	No metadata is provided	Scenarios are provided within the methodology regarding techniques to use during an interview to ensure data quality
			Are data quality controls discussed within the metadata?	The Pilot Study preceded the Census roll out and was followed by a post enumeration survey to test the results	Best practices are discussed in general. Physical implementation of such practices is not clearly expressed in the metadata.	No metadata is provided	Methodology discusses the validation and constraints applied during data capture
		Does the metadata comprehensively describe the data collection methodology?	Are the statistical techniques employed suitably documented?	Effort is made to target all households within the country on a particular day but not documentation of the statistical error achieved is reported on	The 5% data sample test is documented. Individual districts reports are published per year	No metadata is provided	The theoretical development of the weight calculation is provided within the methodology
			Are all data sources used documented?	Not applicable	Data is sourced per school	No metadata is provided	Not applicable
			Is the sampling selection framework and decisions taken adequately documented?	The Ministry of Home Affairs uses a Sample Registration System but does not provide documentation regarding how the system is used in the Census	All schools are targeted, 5% of schools are checked for correctness in the sample survey	No metadata is provided	The sampling plan and the expansion of the sample are detailed within the methodology
			Are the techniques for ensuring anonymity provided?	Data is aggregated to District level ensuring anonymity	Data is published at a district level, not a school. Therefore, surveyed individual's information is protected	No metadata is provided	Data is aggregated over individual households when released
			Is the questionnaire design documented?	The questionnaire structure is not discussed within the metadata	The questionnaire structure is not provided	No metadata is provided	The questionnaire structure is published
			Are the norms and standards regarding the data collection discussed within the metadata?	No references to international norms and standards are made	The data collection methodology is not discussed	No metadata is provided	The Demographic and Health Survey Programme sets the international norms and this is discussed within the Methodology.
			Is the scope of the data collection documented?	The scope of the dataset is not specified	The scope of the survey is not discussed	No metadata is provided	The in an out of scope sections is outlined in the methodology
		Is there an accompanying findings report?	Is a data output report provided together with the dataset publication?	A detailed findings report is provided	Findings reports are provided for each phase of schooling	No specific findings report is provided apart from a published set of statistics from MOSPI about general performance of the country	A detailed findings report is provided

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
		Is the metadata clear and understandable to the data user?	Does the metadata concisely and comprehensively describe how the dataset is produced?	Concepts and definitions are discussed in detail in the Census results documentation in simple language	Concepts and definitions used are understandable, but no analysis of statistical tables are offered to add insight to the reported numbers	No metadata is provided	Concepts and definitions used are discussed in the results documentation and are understandable
				The metadata and final report describes all the available data	The website providing the information could be improved to organise the manner in which the metadata and data is provided. No clear structure to find data and documentation	No metadata is provided	The documentation excludes details about the structure of the data structures
				The results documentation includes references to sources	The documents do not refer to external documentation apart from the provided statistics	No metadata is provided	Each section in the methodology carries references
				Results and metadata accompany the statistical release	Documentation is directly applicable to the data	No metadata is provided	Documentation is directly applicable to the data
	Does the metadata explain the structure of the dataset?	Does the metadata document the structure of the dataset?	Is the physical layout (data tables, fields, database) of the data structures documented?	The physical structures are documented on the Census Website	Table and field structures are not documented	No metadata is provided	The table structures are not detailed
			Are the hardware and software requirements detailed regarding use of the data?	The software and hardware requirements are excluded	Software and hardware requirements are not discussed	No metadata is provided	Software and hardware requirements are not discussed
			Does the metadata explain how to navigate and use online databases if relevant to the data publication	Not applicable - An interactive database tools is not utilised to extract the data	An interactive database tools is not utilised to extract the data	No metadata is provided	An interactive database tools is not utilised to extract the data
Is the data comprehensive for the data user's requirements?	Does the dataset contain all required statistical units?	Are all expected statistical units populated?	Is each combination of statistical units represented within the data as per the required scope of the dataset?	The datasets release provide only a static representation of the data collected per district's spreadsheet. If additional variables are required in a spreadsheet for analysis, a request must be made for such data. Detailed data analysis is only available at a University Workstation, with users not allowed to extract the information apart from physical printouts	The dataset matches the metadata	No metadata is provided	The dataset matches the metadata
		Does the dataset structure match the metadata outline?	Does the dataset layout match the metadata in terms of the table, field names provided?	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided	The dataset matches the metadata
		Does the dataset reasonably convey the	Do values provided within a field broadly match the definition?	The dataset matches the metadata	The dataset matches the metadata	No definitions are provided	The dataset matches the metadata

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
		definitions, scope, classification, valuation and timeline as specified within the metadata?	Are the provided Statistical units matching the scope of the dataset?	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided	The dataset matches the metadata
			Do the categories within the data match the specified classification?	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided	The dataset matches the metadata
			Does the data match the data type of the particular fields?	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided	The dataset matches the metadata
			Does the time periods within the data match the expected time period?	The dataset matches the metadata	The dataset matches the metadata	No metadata is provided	The dataset matches the metadata
	Are all data rules met within the dataset?	Are there blank data entries where the data rule mandates an entry?	As per the data rules, are all necessary fields provided and populated?	All necessary values are provided	All necessary values are provided	The data extract is populated; however it can't be determined what the rules may be	All necessary values are provided
		Are all mandatory attributes populated?	Are Mandatory fields populated?	All mandatory fields are populated	All mandatory fields are populated	There are no blank entries in the statistical reports	All mandatory fields are populated
		Are all inapplicable attributes left blank?	Are inapplicable attributes blank where appropriate	Inapplicable fields are empty	Inapplicable fields are empty	unable to determine which are inapplicable fields due to the lack of metadata	Inapplicable fields are empty
Is the data accurate for the data user's purpose?	Are the data coverage methods adequate?	Has the dataset adequately applied data sampling techniques?	Are the standard error, the coefficient of variation, the confidence interval and the mean square error calculation in line with international norms and standards	Data sampling techniques are not discussed within the metadata or results documentation	Only the 5% sample of schools are tested for accuracy. Greater effort could be made to improve the testing of school data	No metadata is provided	The data sampling calculations are internationally accepted
		Are imputation techniques adequately applied	Is the response rate too low and is imputation rate too high?	The imputation techniques are not discussed within the metadata	Imputation techniques are not discussed within the metadata	No metadata is provided	A 95% confidence interval was achieved across the survey in line with international norms
		Has the dataset adequately applied non sampling techniques	Is the frame coverage, duplication in the frame, number of statistical units out of scope, misclassification errors, measurement errors, processing errors and model	Non sampling calculations are not discussed within the metadata	The rate of discrepancy is referred to in the 5% Sample	No metadata is provided	Frame coverage for India is not discussed. Unclear whether the sample is representative

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
			assumption errors data calculations in line with international norms and standards?	🟥	🟩	🟥	🟥
	Is the data suitable for reporting?	Does the provided data correspond with a comparative source?	The aggregated data of the dataset equates to the comparative data source	A pilot study and post enumeration study were conducted. The calculated coverage errors were not published	The data collection is compared per school only against schools from the 5% sample per district	No comparative source or pilot study is identified.	The survey compares suitably to the pilot study
		Has a downstream source been quality assessed?	Was the primary data source assessed in line with the Public Data Quality Assessment Framework	Not applicable	Not applicable	Not applicable, cannot be determined	Not applicable
		Are the adopted maintenance procedures suitable?	Has the sample frame been regularly updated and managed?	The sampling frame is not discussed within the metadata	All schools are targeted; therefore, no sample frame is used	No references to a sample frame are provided	According to the methodology, the sample frame was established on initiation of the project
			Does the organisation conduct regularly quality assurance procedures?	The quality assurance practices are not documented	There are no data correction practices in place even where discrepancies are found amongst the 5% sample	No metadata is provided to make a determination	The survey is not regularly collected
			Does the organisation conduct regular data audits and are the errors identified acceptable?	The Office of the Registrar General is audited by the Auditor General	NUEPA is regularly audited by the Auditor General of India	National Crime Records Burea is regularly audited	The DHS Programme is not audited as governments are
Is the data clearly understandable to the data user?	Is there consistency between terms used and the naming convention?	Is there consistency between the terms used and the organisation's definitions?	Are concepts, definitions, classifications and standards applicable to the dataset made available?	Concepts are consistently applied across the Census. Where there are changes to boundaries or question options they are discussed within the metadata	Terms are consistently used across the different tables and datasets	There is no metadata to determine if the definitions are employed correctly	Concepts are consistently applied
			Are the terms used based on agreed company definitions?	Terms used are based on concepts in the metadata	Definitions follow those provided in the guidebook across datasets	There is no metadata to determine if the definitions are employed correctly	Terms follow Demographic and Health Survey Programme Prescripts
Is the data applicable to the data user?	Does the dataset meet the needs of the data users?	Does the dataset meet the requirements of the data users?	Have the public's data requirements been determined and have they been made available?	The public's data requirements identified in the pilot study are not discussed	The public's data requirements have not been determined or are not made available	The public's data requirements have not been determined or are not made available	Cannot be determined if all requirements of the public are captured within the Census
			Do the identified data requirements	The public's data requirements identified in the pilot study are not discussed	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
			correspond with the purpose of the dataset?				
			Is the public satisfied with the contents of the dataset?	The public's data requirements identified in the pilot study are not discussed	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented
	Does the dataset contain beneficial information for the data user?	Does the dataset contain unnecessary/superfluous details not required by the data user?	All sub-components of the dataset must be examined to determine the granular subunits are relevant to the data user	Technical information and unimportant data has been excluded	Technical information and unimportant data has been excluded	Reported data is pertinent to crime reporting	Technical information and unimportant data has been excluded
Is the data concise enough for the data user?	Is the data concisely presented to the public?	Is the data to the point?	Does the data contain unnecessary elements for the data user?	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant
			Does the data provide a sufficient level of granularity for the data user?	Information lower than district may be required amongst some communities. This is currently not freely available.	School level information is not provided, data aggregated to a district	Data users may require greater granularity per thematic area or geographical location	Individual responses are provided
Is the data consistently represented to the data user?	Is the data semantically consistent?	Are data rules and definitions applied consistently to the dataset?	Does the data provider follow a common methodology (international standard) when producing the dataset?	No standards are referenced in the metadata	The data collection methodology is not discussed	The data collection methodology is not discussed	The questionnaire construction is consistent with Demographic and Health Survey Programme
			Does the metadata describe changes to the methodology and/or rules, definitions that are used?	The metadata describes changes to the terms used between census 2001 and 2011	The data collection methodology is not discussed	The data collection methodology is not discussed	There were no changes made to interpretations of terms
	Is the data structurally consistent?	Are data values consistently reported to the public?	Are data values over time are reported consistently?	The data matches the results documentation	The data matches the results documentation	Data is consistently reported over the period of available data	The data matches the results documentation
			Does the data provider reasons for changes in data values?	There were no changes made to the interpretations of terms	There were no changes made to interpretations of terms	If there were changes, it is not documented and therefore one is unaware if there were issues	There were no changes made to interpretations of terms
Is the data current enough and released in a timely manner for the data user?	Is the dataset punctually released?	Is the dataset released in a punctual manner?	Is the average time between the end of the data collection and the data release within recommended time-frames?	The data release date of the census is not published	The data is released annually with a 2 year delay	Crime statistics are released two years after collation annually	The time taken to publish the survey after completion of data collection activities is not published
			Does the organisation release a report on the time frame for data releases?	Not applicable - The census is carried out every 10 ten years	A time frame for reporting is not provided	A time frame for reporting is not provided	The survey is carried out every 10 ten years

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
Is the data accessible to the data user?	Are data access controls too stringent?	Are the applicable data users granted an appropriate level of data access?	Are the access rights clearly and uniformly applied ensuring that the correct individuals are granted access?	Where greater detail or customised tabulation is required, a request must be sent to the Office of the Registrar explaining how the data would be used. Where access to microdata is required, the user needs to find a University Workstation. Data downloads are not permitted	Freely available, limited provision of detailed data	Freely available - no limitations, detailed information is unavailable	Data access is on request only via the DHS Website, following the request protocol
			Can users easily find directions on how to access the required information?	Static files are easily accessible, detailed data is difficult to work with and access	Information is available on the DISE Website; however, it is not well organised. Most reports are provided in PDF format. A single large extract is provided in Excel covering all data themes	No supporting documentation is provided	Directions are provided on the DHS website
			Does the data provider respond to data requests in line with international standards on responsiveness	No response was received from the provider, when a request was made.	Not applicable - Data is available for download only	Data is available for download; no additional user support is provided for more detailed information	Response are timely
	Is a suitable granularity of data accessible to the public?	Are access controls applied to the correct fields and categories?	Are access controls applied to the correct selection of fields and or categories?	Access to customised tabulation is limited, which limits analysis options	Only aggregated information is released. School level data is not published	Granular data is not made available, there are no mechanisms to request greater detail.	Data access is on request only via the DHS Website
			Adequate levels of anonymity need to be applied whilst ensuring that the user's needs for data are considered	Data has been anonymised	Names of schools, learners are removed	Data is aggregated to municipal level ensuring anonymity	Names of respondents are removed
			Does the metadata describe the levels of detail that are accessible?	The results documentation discusses the detail available within the Census, supporting codes and descriptions are available for download	The metadata does not discuss the table structures that are provided by DISE	No metadata is provided	The metadata does not describe the geographic details or the file structures
	Is the data and metadata provided in a meaningful and useful manner?	Is data and metadata provided in a useful and meaningful manner?	Is the data dissemination strategy shared with the public?	A data dissemination strategy document is not compiled. A brief note is provided on the website stating how further data can be accessed	There is no published data dissemination strategy, but the data is released publicly	There is no published data dissemination strategy, but the data is released publicly	There is no published data dissemination strategy
			Is the data release scheduled?	The next Census is not dated	The next release is not dated	The next release is not dated	The next release is not dated
			Does the metadata provide guidance on how to access and retrieve data using the available mediums?	Tables are provided without guidance on how to use the various mediums available	The metadata does not provide steps on how to access or retrieve data	No metadata is provided	Tables are provided without guidance on how to use the various mediums available

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
			Is the data dissemination strategy frequently updated?	Not applicable	Not applicable	Not applicable	Not applicable
			Does the data dissemination strategy take heed of the needs of the data users?	Database tools could be adopted to improve usability when interacting the data. Data is primarily made available by static files	Not applicable - Data is published with no dissemination strategy documented	There is no data dissemination strategy that is published	Database tools could be adopted to improve usability when interacting the data. Data is primarily made available by static files
	Is the channel of data delivery appropriate for the data user?	Is the channel of data delivery suitable for the needs of the user?	Are hardware and software requirements clearly communicated to the data user?	Software and hardware requirements are not discussed	Software and hardware requirements are not explicitly discussed	Software and hardware requirements are not explicitly discussed	Software and hardware requirements are not discussed
			Have the user's preferred means of accessing data been taken into consideration when developing the data access channel?	Data users are not considered	The channel could be improved with database tools, the data provided by PDF is difficult to work with. Data in excel is only sparsely provided	The channel could be improved with database tools, but Excel is a common preference for distributing data. Accessing data via the MOSPI website is cumbersome	The channel could be improved
			Are the disseminated datasets accessible in multiple data formats?	Data is only available in Excel; detailed data is limited	Provided only in PDF and Excel	Provided only in Excel	Multiple data formats provided
			Where database tools are provided, is the mode of access understandable and simple to operate?	Not applicable	Not applicable	Not applicable	Not applicable
			Are user manuals provided to aid data access to database tools?	Not applicable	Not applicable	Not applicable	Not applicable
			Where external staff resources need to be contacted to request dataset, are such staff reachable?	Not applicable	Not applicable	Not applicable	Staff manning data requests respond speedily
Does the data have integrity when consumed by the data user?	Are data quality constraints suitable for the needs of the data user?	Do the data quality constraints applied within the data system sufficient to manage data integrity?	Do the data quality constraints follow international best practices?	Data constraints are not discussed within the metadata	Data quality constraints are not discussed	Data quality constraints are not discussed	Data input constraints are applied to the electronic form capture used in the survey
			Are the data rules suitably captured within the data quality constraints?	Data constraints are not discussed within the metadata	Data quality constraints are not discussed	Data quality constraints are not discussed	Only data values as discussed within the metadata are provided
			Is the process to address data quality breaches discussed within the metadata?	The process for resolving data quality breaches are not detailed within the metadata	Data quality constraints are not discussed	Data quality constraints are not discussed	The process for resolving data quality breaches are not detailed within the Methodology

Dimension	Element	Indicator	Focal Issue	Census of India 2011	DISE 2014-15	MOSPI - Crime Statistics	DHS 2005/06
	Does the data provider introduce the most appropriate statistical and methodological approaches?	Are the data transformation techniques suitable?	Are the applied data transformation techniques applied within the bounds of international best practices?	The data aggregations do not follow any particular international standard	The data transformation techniques are not discussed	There is no metadata to assess the data transformation techniques applied	The data provided is not aggregated
			Are the data transformation methodologies detailed within the supporting metadata?	Aggregation matches the metadata	The data transformation techniques are not discussed	There is no metadata to assess the data transformation techniques applied	The data provided is not aggregated
Is the data traceable?	Does the data provider track the data source?	Does the organisation track data transformations ?	Does the organisation document the data sources and transformations which are used within the metadata?	Not applicable - no source systems required	Not applicable - no source systems required	There is no metadata to assess the quality of the source systems that are used	Not applicable - no source systems required
			Has supporting information been vetted using the public data quality assessment framework?	Not applicable - no source systems required	Not applicable - no source systems required	There is no metadata to assess the quality of the source systems that are used	Not applicable - no source systems required

Appendix 5

Data Quality Assessment of South African Datasets

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
Is the Metadata a useful to the data user?	Does the metadata a convey the contents of the dataset?	Are concepts and definitions and classification provided within the metadata to describe the underlying data?	Are definitions made available within the metadata?	Concepts and definitions are discussed in detail in the Census metadata	Concepts and definitions are discussed in detail in the Community Survey metadata	Concepts and definitions are discussed in detail in the GHS metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Definitions are not discussed within the ANA report	Definitions are provided within the Final report	Terms and definitions are not explicitly defined	Definitions of terms are provided in an accompanying codebook of each wave of the study	Definitions of terms are provided in an accompanying codebook the survey	A codebook is provided without contextual definitions describing each term in the questionnaire apart from options available within the questionnaire
			Are deviations from standards reported?	All concepts are well defined throughout the various pieces of documentation	All concepts are well defined throughout the various pieces of documentation	All concepts are well defined throughout the various pieces of documentation	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No standards are identified	References are not made to international standards within the documentations	Due to the lack of documentation regarding definitions, it cannot be determined if the definitions follow international standards	All concepts are well defined throughout the various pieces of documentation	All concepts are well defined throughout the various pieces of documentation	Concepts are weakly defined
		Is the metadata up to date?	Is a document register regularly maintained in line with the data collection strategy?	The documentation is provided for each census	The documentation directly describes CS2016	The documentation directly describes GHS2015	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	A data register is not provided	The documentation is provided for the particular survey alone	The documentation is included in the results report of the survey	The documentation is provided for each wave of the study	The documentation follows the rollout of the survey results	The documentation follows the rollout of the survey results
		Are all tables and fields defined?	Is the purpose and definition of each table and field within	Each data file is discussed in detail in the Metadata Annexures	Each data file is discussed in detail in the Metadata Annexures	Each data file is discussed in detail in the Metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Tables and fields are not defined	Fields are not described within the data extract nor Individual table structures	Table and field details are not provided within the documentation	Each data file is discussed in detail in the Codebook	Each data file is discussed in detail in metadata	Each data file structure is discussed within the codebook

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			the dataset reported on?	Fields used in the census are consistent with Stats SA's other products	Fields used in the Community Survey are consistent with Stats SA's other products	Fields used in the GHS are consistent with Stats SA's other products	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Tables and fields are not defined	Not applicable - Fields are used are specific to Youth Risk and cannot be compared to other surveys	Data is only released within the Survey Report. The fields within the dataset are not clearly discussed	Fields are consistent across waves	Fields are consistently applied across all countries running the survey	Fields are consistently applied across the surveys caried out each year
		Is the geographic distribution clearly defined?	Is the geographic level of data granularity described within the metadata ?	The geographic granularity is well documented within the metadata	The geographic granularity is well documented within the metadata	The geographic granularity is well documented within the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The education geography dimension is not described	The Geographic granularity of the data output is not detailed within the documentation - although the findings report mentions that the data is provided at Provincial level	The Geographic granularity of the data output is not detailed within the documentation	Not applicable: - The data can only be reported at National Level	Geography is defined within the survey stratification to Province level	The Geographic structure is referred but not discussed in detail
				Geographic boundaries are well documented within the metadata	Geographic boundaries are well documented within the metadata	Geographic boundaries are well documented within the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The education geography dimension is not described	Geographic boundaries are not well documented - detail is not required in great detail	Geographic boundaries are not documented	Not applicable: - The data can only be reported at National Level	Not applicable: Changes to boundaries are not discussed due to provincial boundary consistency	Not applicable: Changes to boundaries are not discussed due to provincial boundary consistency
	Does the metadata describe the context of the dataset?	Does the metadata describe the data quality practices applied when producing the dataset?	Does the organisation define data quality within the metadata ?	The methodology documentation discusses the validation processes that are implement and refers to the SASQAF principles	The methodology documentation discusses the validation processes that are implement and refers to the SASQAF principles	The methodology documentation discusses the validation processes that are implement and refers to the SASQAF principles	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No definition of data quality is provided	The methodology documentation discusses the sampling plan, sample structure, the results of a Pilot Study and the weighting calculations used	Data quality is discussed within the Appendices of the Survey Report	Data quality is not defined explicitly, but the technical documentation outlines the various processes implemented to ensure data quality is maintained	Data quality is not defined explicitly, but the technical documentation outlines the various processes implemented to ensure data quality is maintained	Data quality processes are not discussed within the user guide

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			How does the organisation measure data quality within the metadata?	The metadata does not measure data quality but applies the SASQAF principles	The metadata does not measure data quality but applies the SASQAF principles	The metadata does not measure data quality but applies the SASQAF principles	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	A verification process is discussed using a Pilot study	The metadata is not explicit about how data quality is measured but sites a 95% confidence interval of reported statistcs when compared to the pilot study	The document reports on response rates, confidence internvals and sampling errors as a proxy for data quality	Response rates are provided with each data release which are generally above 90% in all instances	Although data quality is not measured directly the metadata notes that the survey achieved a 95% confidence level	Data quality processes are not discussed within the user guide
			How does the organisation process data concerns? Is this discussed in the metadata?	The metadata discusses the categories of errors that emerged and the processes that were put in place to address these issues.	The metadata discusses the categories of errors that emerged and the processes that were put in place to address these issues.	The metadata discusses the categories of errors that emerged and the processes that were put in place to address these issues.	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Internal and unpublished processes within the department	Error handling is not discussed within the final report	The documentation does not describe how data faults are address	Data quality controllers are highered and field workers are sent back into the field where questionnaires have serious concerns. In addition, call back confirmations are employed to ensure the right households are targeted	The quality assurance procedures are detailed in Operations for data collection such as contacting schools and scoring their responses	Quality assurance processes are not discussed
			Are data quality controls discussed within the metadata?	The statistical release details the controls that are put in place when managing the census	The statistical release details the controls that are put in place when managing the CS	The statistical release details the controls that are put in place when managing the GHS	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Data Quality controls are not discussed within the metadata	Data validation process are discussed only	Training was provided to field workers regarding the correct use of the questionnaire and extra checks were in place during data capture from paper to the database	The statistical release details the controls that are put in place when managing the census	Quality assurance is included as part of the data collection process	Quality assurance processes are not discussed

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
		Does the metadata comprehensively describe the data collection methodology?	Are the statistical techniques employed suitably documented?	Detail is provided within the documentation detailing how each enumeration area is identied the processes employed to ensure the right targets are reached	The Sampling frame and Enuermation Area Sample size and coverage are discussed in the technical report	The Sampling frame and Enuermation Area Sample size and coverage are discussed in the technical report	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Statistical techniques are not discussed	Data Sampling and the conducting of the pilot study follows international best practices	Sampling errors, response rates, confidence intervals amongst other factors are thoroughly discussed	The statistical checks are documented in detail in the technical documentation	The statistical checks are documented in detail in the technical documentation	The sampling frame is referred to but detailed information regarding the process is not communicated within the metadata
			Are all data sources used documented?	Not applicable	Not applicable	Not applicable	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
			Is the sampling selection framework and decisions taken adequately documented?	Decisions regarding how the enumeration areas are targeted and covered are discussed	Decisions regarding how the sample frame was targeted and covered in the survey are discussed	Decisions regarding how the sample frame was targeted and covered in the survey are discussed	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The sampling framework is not adequately provided	The sampling plan and the expansion of the sample are detailed within the methodology	The sampling plan and the expansion of the sample are detailed within the survey report	The Sampling Frame is based on Statistics South Africa's Master Sample and the structure is documented in the technical documents	The sampling framework and process is comprehensively documented	The sampling frame is referred to but detailed information regarding the process is not communicated within the metadata
			Are the techniques for ensuring anonymity provided?	Data is aggregated to a Sub Place level when using Super-cross and small place on request. Individual response is protected from the public	Data is aggregated to a Municipality due to representivity concerns when released, therefore individual details are protected	Data is aggregated to a Province/Metro due to representivity concerns when released, therefore individual details are protected	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Detailed data is not provided	Data is aggregated to provincial level when released within the documentation only	Data is aggregated in the survey report. Data is very aggregated	Data is aggregated to National level is annonymised over the individual	Individual details are removed from the data provided	Individual details are removed from the data provided
			Is the questionnaire design documented?	The questionnaire structure is well documented	The questionnaire structure is well documented	The questionnaire structure is well documented	No metadata is provided	No metadata is provided	The questionnaire is published	No metadata is provided	Details about the questionnaire is not published	The questionnaire structure is published	The questionnaire structure is published as an appendix	The questionnaire structure is well documented	The questionnaire structure is well documented	The questionnaire structure is documented in the codebook and the questionnaire is published

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
				🟩	🟩	🟩	🟥	🟥	🟨	🟥	🟥	🟩	to the survey report	🟩	🟩	🟩
			Are the norms and standards regarding the data collection discussed within the metadata?	The structure and standards implemented within the Census are guided by international experts and are documented	The structure and standards implemented within the Community Survey are guided by international experts and are documented	The structure and standards implemented within the GHS are guided by international experts and are documented	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No norms were identified	International norms are not referenced	The documents do not specific if the DHS is linked to the international DHS Programme as in India and Brazil	Data collection practices are guided by those adopted by Stats SA but not clearly documented within the metadadta	Data collection process follow internationally accepted TIMMS processes	The data collection process is not discussed in detail
			Is the scope of the data collection documented?	The scope and out of scope data is outlined in the metadata	The scope and out of scope data is outlined in the metadata	The scope and out of scope data is outlined in the statistical release	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The scope is not discussed	The scope is not discussed within the documentation	The scope is not discussed within the methodology	The Out of scope dataset are outlined in the metadata	The scope of the study is not clearly presented, but the focus of the survey clearly follows Math and Science trends	The scope of the study is not discussed
		Is there an accompanying findings report?	Is a data output report provided together with the dataset publication?	A detailed findings report is provided	A detailed findings report is provided	A detailed findings report is provided	No findings report is provided apart from the statistics	No findings report is provided apart from the statistics	No findings report is provided apart from the statistics in the Education at a Glance release	A NEIMS reports is published with screenshots of data from the NEIMS data collection	A findings report is produced with limited i analysis	A detailed findings report is provided	The survey findings report is very detailed	A detailed findings report is provided	A detailed findings report is provided	A detailed findings report is provided
		Is the metadata clear and understandable to the data user?	Does the metadata concisely and comprehensively describe how the dataset is produced?	Concepts and definitions are discussed in detail in the Census results documentation in simple language	Concepts and definitions are discussed in detail in the Community Survey results documentation in simple language	Concepts and definitions are discussed in detail in the GHS results documentation in simple language	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Concepts and definitions used are discussed in the results documentation and are understandable	Definitions are not provided	Concepts and definitions are discussed clearly for each data release	Concepts and definitions are discussed clearly for each data release	Concepts are not explained only presented without context
				The sequence of information presented in the documentation	The sequence of information presented in the documentation	The sequence of information presented in the documentation	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The documentation excludes details about the structure	The documentation excludes details about the structure	The sequence of information presented in the documentation	The sequence of information presented in the documentation	The sequence of information presented in the documentation is provided well

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
				n is provided well	n is provided well	n is provided well						of the data structures	of the data structures	n is provided well	n is provided well	
				The statistical techniques are described and referenced well	The statistical techniques are described and referenced well	The statistical techniques are described and referenced well	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Limited or no references are made to data	The documentation is referenced	The statistical techniques are described and referenced well	The statistical techniques are described and referenced well	the User guide is not referenced
				Documentation is directly applicable to the data	Documentation is directly applicable to the data	Documentation is directly applicable to the data	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Documentation is directly applicable to the data	Documentation is directly applicable to the data	Documentation is directly applicable to the data	Documentation is directly applicable to the data	Documentation is directly applicable to the data
	Does the metadata explain the structure of the dataset?	Does the metadata document the structure of the dataset?	Is the physical layout (data tables, fields, database) of the data structures documented?	The structure of the data files that are disseminated are well documented	The detailed data is not ready for publication	The structure of the data files that are disseminated are well documented	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The table structures are not detailed	The table structures are not detailed	The structure of the data files that are disseminated are well documented	The structure of the data files that are disseminated are well documented	The structure of the data files that are disseminated are well documented
			Are the hardware and software requirements detailed regarding use of the data?	Software requirements are discussed within the metadata	Not applicable - data not released yet	Software requirements are discussed within the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Software and hardware requirements are not discussed	Software and hardware requirements are not discussed	Software requirements are discussed within the metadata	Software requirements are discussed within the metadata	Software requirements are briefly discussed within the metadata
			Does the metadata explain how to navigate and use online databases if relevant to the data publication	A handbook is produced to guide users on how to access the various database tools	Not applicable - data not released yet	Documentation on using the Nesstar tool is not provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	An interactive database tools is not utilised to extract the data	Not applicable - An interactive database tools is not utilised to extract the data	A handbook is produced to guide users on how to access the various database tools	Not applicable - HSRC does not provide online database tools	Not applicable - HSRC does not provide online database tools

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
Is the data comprehensive for the data user's requirements?	Does the dataset contain all required statistical units?	Are all expected statistical units populated?	Is each combination of statistical units represented within the data as per the required scope of the dataset?	The dataset matches the metadata	Not applicable - data only available in statistical report	The dataset matches the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Detailed data is not provided	The metadata does not describe the fields and tables, cannot be assessed	A dataset is not published	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata
		Does the dataset structure match the metadata outline?	Does the dataset layout match the metadata in terms of the table, field names provided ?	The dataset matches the metadata	Not applicable - data only available in statistical report	The dataset matches the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Detailed data is not provided	The metadata does not describe the fields and tables, cannot be assessed	A dataset is not published	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata
		Does the dataset reasonably convey the definitions, scope, classification , valuation and timeline as specified within the metadata?	Do values provided within a field broadly match the definition?	The dataset matches the metadata	Not applicable - data only available in statistical report	The dataset matches the metadata	No definitions are provided	No definitions are provided	No definitions are provided	No definitions are provided	Detailed data is not provided	The metadata does not describe the fields and tables, cannot be assessed	A dataset is not published	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata
			Are the provided Statistical units matching the scope of the dataset?	The dataset matches the metadata	Not applicable - data only available in statistical report	The dataset matches the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Detailed data is not provided	The metadata does not describe the fields and tables, cannot be assessed	A dataset is not published	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata
			Do the categories within the data match the specified classification?	The dataset matches the metadata	Not applicable - data only available in statistical report	The dataset matches the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Detailed data is not provided	The metadata does not describe the fields and tables, cannot be assessed	A dataset is not published	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			Does the data match the data type of the particular fields?	The dataset matches the metadata	Not applicable - data only available in statistical report	The dataset matches the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Detailed data is not provided	The metadata does not describe the fields and tables, cannot be assessed	A dataset is not published	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata
			Does the time periods within the data match the expected time period?	The dataset matches the metadata	Not applicable - data only available in statistical report	The dataset matches the metadata	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Detailed data is not provided	The metadata does not describe the fields and tables, cannot be assessed	The reported data follows the reported time period	The dataset matches the metadata	The dataset matches the metadata	The dataset matches the metadata
	Are all data rules met within the dataset?	Are there blank data entries where the data rule mandates an entry?	As per the data rules, are all necessary fields provided and populated?	All necessary values are provided	Not applicable - cannot be determined as yet	All necessary values are provided	The data extract is populated; however, it can't be determined what the rules may be	The data extract is populated; however, it can't be determined what the rules may be	Detailed statistics and metadata is not released to make a determination	Detailed statistics and metadata is not released to make a determination	Detailed statistics and metadata is not released to make a determination	All necessary values are provided as per the questionnaire	A dataset is not published	All necessary values are provided	All necessary values are provided	The data rules are not discussed in detail within the metadata
		Are all mandatory attributes populated?	Are Mandatory fields populated?	All mandatory fields are populated	Not applicable - cannot be determined as yet	All mandatory fields are populated	There are no blank entries in the statistical reports	There are some blank entries in the statistical reports	There are some blank entries in the statistical reports	Detailed statistics and metadata is not released to make a determination	Detailed statistics and metadata is not released to make a determination	All necessary values are provided as per the questionnaire	A dataset is not published	All mandatory fields are populated	All mandatory fields are populated	All mandatory fields as per the questionnaire instructions are populated
		Are all inapplicable attributes left blank?	Are inapplicable attributes blank where appropriate	Inapplicable fields are empty	Not applicable - cannot be determined as yet	Inapplicable fields are empty	unable to determine which are inapplicable fields due to the lack of metadata	unable to determine which are inapplicable fields due to the lack of metadata	unable to determine which are inapplicable fields due to the lack of metadata	Detailed statistics and metadata is not released to make a determination	Detailed statistics and metadata is not released to make a determination	All necessary values are provided as per the questionnaire	A dataset is not published	Inapplicable fields are empty	Inapplicable fields are empty	Inapplicable fields are empty
Is the data accurate for the data user's purpose?	Are the data coverage methods adequate?	Has the dataset adequately applied data sampling techniques?	Are the standard error, the coefficient of variation, the confidence interval and the	The manner the enumeration area was targeted has been internationally accepted by the council of experts	The manner the enumeration area was targeted has been internationally accepted by the council of experts	The manner the enumeration area was targeted has been internationally accepted by the council of experts	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The data sampling calculations follow international best practices	The data sampling calculations are internationally accepted	The response rate and confidence intervals are well within general international norms	The response rate and confidence intervals are well within general international norms	Data sampling calculations are not provided

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			mean square error calculation in line with international norms and standards	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟥
		Are imputation techniques adequately applied	Is the response rate too low and is imputation rate too high?	Improved techniques were used to reduce the need for imputation using electronic questionnaires as well as post enumeration survey that was used to verify 20% of the census enumeration areas. These results of these findings were accepted and documented	Imputation only applied to extreme errors. Logical imputation was applied mostly. The response rate is within standards. The imputation rate is not provided	The response rate is equal or above the international standard	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	A 95% confidence interval was achieved when comparing data to the pilot survey in line with international norms	The relative error rates are within accepted ranges	The response rate is above 90% and imputation is avoided but the imputation rate is not discussed within the documentation	The confidence interval is above 95% and imputation is avoided but the imputation rate is not discussed within the documentation	Data sampling calculations are not provided
		Has the dataset adequately applied non sampling techniques	Is the frame coverage, duplication in the frame, number of statistical units out of scope, misclassification errors, measurement errors, processin	The enumeration area covered all households in South Africa	The sampling frame is within international standards	Non sampling calculations are within international norms and standards, using the Statistics South Africa's Household Master Sample Frame based on Census 2011 enumeration areas	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The frame sample is based on the school level data from the department for secondary schools. The accuracy of the sample framework is not discussed	The confidence intervals are within accepted ranges	The survey only attempted to be Nationally representative by design	The survey only attempted to be Provincially representative by design	The survey only attempted to be Provincially representative by design

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			g errors and model assumption errors data calculations in line with international norms and standards ?	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟨	🟩	🟨	🟩	🟩
	Is the data suitable for reporting?	Does the provided data correspond with a comparative source?	The aggregated data of the dataset equates to the comparative data source	The census results compared adequately with the Census Pilot Survey and Post Enumeration Survey	The CS results compared adequately with the Pilot Survey	The GHS results compared to the previous GHS survey for consistency, the data values cover are not compared against an alternate source	No comparative source or pilot study is identified.	No comparative source or pilot study is identified.	No comparative source or pilot study is identified.	No comparative source or pilot study is identified.	A pilot study is identified but no results are published of the comparison	The survey compares suitably to the pilot study	No comparative source is identified	A comparative source is not discussed within the metadata	A comparative source is not discussed within the metadata	A comparative source is not discussed within the metadata
		Has a downstream source been quality assessed?	Was the primary data source assessed in line with the Public Data Quality Assessment Framework	Not applicable	Not applicable	Not applicable	Not applicable	Data collections are not referenced, unsure of how the list is produced	Data collections are not referenced, unsure of how the list is produced	Data collections are not referenced, unsure of how the list is produced	Data collections are not referenced, unsure of how the list is produced	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable
		Are the adopted maintenance procedures suitable?	Has the sample frame been regularly updated and managed ?	According to the methodology, the enumeration areas are reassessed at the start of the census	According to the methodology, the enumeration areas are reassessed at the start of the CS	According to the methodology, the enumeration areas are reassessed at the start of the survey but follow largely from	No references to a sample frame are provided	No references to a sample frame are provided	No references to a sample frame are provided	No references to a sample frame are provided	No references to a sample frame are provided	According to the methodology, the sample frame was established on initiation of the project	According to the methodology, the sample frame was established on initiation of the project	According to the methodology, the enumeration areas are reassessed at the start of the census	The sample frame is reassessed before the study and is thoroughly documented	The sample frame is reassessed before the study and is thoroughly documented

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
						the Census 2011 enumeration areas										
			Does the organisation conduct regularly quality assurance procedures?	Each survey applies the SASQAF	Each survey applies the SASQAF	Each survey applies the SASQAF	No metadata is provided to make a determination	No metadata is provided to make a determination	No metadata is provided to make a determination	No metadata is provided to make a determination	No metadata is provided to make a determination	The survey is not regularly collected	The survey is not regularly collected	Data quality checks are built into the process managed by quality supervisors	Data quality checks are built into the process managed by quality supervisors	Data quality checks are not built into the data collection process or are not documented
			Does the organisation conduct regular data audits and are the errors identified acceptable?	Statistics South Africa is independently audited following South African Public Policy	Statistics South Africa is independently audited following South African Public Policy	Statistics South Africa is independently audited following South African Public Policy	SAPS is audited annually, but performs poorly	DBE is internally audited and by the Auditor General. The results of the data review are not released	DBE is internally audited and by the Auditor General. The results of the data review are not released	DBE is internally audited and by the Auditor General. The results of the data review are not released	DBE is internally audited and by the Auditor General. The results of the data review are not released	The MRC is externally audited, although the YRBS study results are not mentioned within such audits	The MRC is externally audited	UCT has a University Audit Committee monitoring all aspects of the organisation including SALDRU	The HSRC is audited by Public Protector	The HSRC is audited by Public Protector
Is the data clearly understandable to the data user?	Is there consistency between terms used and the naming convention?	Is there consistency between the terms used and the organisation's definitions?	Are concepts, definitions, classifications and standards applicable to the dataset made available?	Concepts are consistently applied across Statistics South Africa Surveys	Concepts are consistently applied across Statistics South Africa Surveys	Concepts are consistently applied across Statistics South Africa Surveys	There is no metadata to determine if the definitions are employed correctly	There is no metadata to determine if the definitions are employed correctly	There is no metadata to determine if the definitions are employed correctly	There is no metadata to determine if the definitions are employed correctly	There is no metadata to determine if the definitions are employed correctly	Concepts are consistently applied	Concepts are not defined within the documentation	Concepts are consistently applied across NIDS Waves	Concepts are consistently applied across TIMSS countries to ensure comparability and the public is informed via published metadata	Concepts are consistently referred to across the annual SASAS survey
			Are the terms used based on agreed company definitions?	Terms follow agreements based on discussions amongst the council of international and South	Terms follow agreements based on discussions amongst the council of international and South	Terms follow agreements based on discussions amongst the council of international and South	There is no metadata to determine if the definitions are employed correctly	There is no metadata to determine if the definitions are employed correctly	There is no metadata to determine if the definitions are employed correctly	There is no metadata to determine if the definitions are employed correctly	There is no metadata to determine if the definitions are employed correctly	Terms follow previous YRBS surveys	Concepts are not defined within the documentation	Terms are consistent with other South African Surveys	Terms follow TIMSS standards	Terms follow SASAS reporting consistently

Dimension	**Element**	**Indicator**	**Focal Issue**	**Census of South Africa 2011**	**Community Survey 2016**	**General Household Survey 2015**	**SA Crime Statistics**	**School Master List**	**Annual School Survey 2014**	**NEIMS 2016**	**Annual National Assessment 2014**	**YRBS 2011**	**Demographic and Health Survey 2003**	**National Income Dynamics Study Wave 1,2,3,4**	**Trends in Mathematics and Science Survey 2011**	**South African Social Attitudes Survey 2012**
				African experts	African experts	African experts	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟩	🟩
Is the data applicable to the data user?	Does the dataset meet the needs of the data users?	Does the dataset meet the requirements of the data users?	Have the public’s data requirements been determined and have they been made available ?	The publics data requirements are detailed within the metadata	The subject area specialists were consulted to determine data requirements of the public and these are briefly discussed within the metadata	The publics data requirements are detailed within the metadata	The public’s data requirements have not been determined or are not made available	The public’s data requirements have not been determined or are not made available	The public’s data requirements have not been determined or are not made available	The public’s data requirements have not been determined or are not made available	The public’s data requirements have not been determined or are not made available	Cannot be determined if all requirements of the public are captured within the Census	Not applicable - Cannot be determined if all requirements of the public are captured within the Census	8 research papers were commissioned to determine the state of need across central national concerns	The TIMSS study follows an international investigation is not based on local needs	The study is not based on user's needs
			Do the identified data requirements correspond with the purpose of the dataset?	The requirements match against the structure of the questionnaires	The requirements match against the structure of the questionnaires	The requirements match against the structure of the questionnaires	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable	Requirements were not assessed - not applicable	Requirements were not assessed - Not applicable	Not applicable - Requirements were not assessed	The questionnaires were designed following the findings of the 8 formative papers	The TIMSS study follows an international investigation is not based on local needs	The study is not based on user's needs
			Is the public satisfied with the contents of the dataset?	The metadata does not discuss whether the public is satisfied with the rollout of the census	The metadata does not discuss whether the public is satisfied with the rollout of the Community Survey	The metadata does not discuss whether the public is satisfied with the rollout of the GHS	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented	The publics opinions of the data release are not documented	The metadata does not discuss whether the public is satisfied with the rollout of the census	The metadata does not discuss whether the public is satisfied with the rollout of the census	The metadata does not discuss whether the public is satisfied with the rollout of the census
	Does the dataset contain beneficial information for the data user?	Does the dataset contain unnecessary/ superfluous details not required by the data user?	All sub-components of the dataset must be examined to determine the granular subunits are relevant	Technical information and unimportant data has been excluded	Technical information and unimportant data has been excluded	Technical information and unimportant data has been excluded	Reported data is pertinent to crime reporting	Reported data is very relevant to school management	Reported data is very relevant to school management	Reported data is very relevant to school management	Reported data is very relevant to school management	Extra weights are included in the detailed data without explanation in the metadata	Cannot be determined as the dataset is not published	Technical information and unimportant data has been excluded	Technical information and unimportant data has been excluded	Technical information and unimportant data has been excluded

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			to the data user	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟨	🟩	🟩	🟩
Is the data concise enough for the data user?	Is the data concisely presented to the public?	Is the data to the point?	Does the data contain unnecessary elements for the data user?	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant	Cannot be determined as the dataset is not published	All fields provided are relevant	All fields provided are relevant	All fields provided are relevant
			Does the data provide a sufficient level of granularity for the data user?	Small Area data is also available on request. Data granularity is suitable	Not applicable - Detailed data is not ready for publication	Small Area data is also available on request. Data granularity is suitable	Data users may require greater granularity per thematic area or geographical location	Granularity is at a school level with GIS codes	Data published is at a Province level only	Data published is at a Province level only	Data published is at a Province level only	Individual responses are provided	Cannot be determined as the dataset is not published	Data is only available at National level but this is also necessary due to the nature of a panel study where respondents can change location at will year after year	Data is available at Province level only	Data is available at Province level only
Is the data consistently represented to the data user?	Is the data semantically consistent?	Are data rules and definitions applied consistently to the dataset?	Does the data provider follow a common methodology (international standard) when producing the dataset?	The questionnaire construction is guided by international experts	The questionnaire construction is guided by international experts	The questionnaire construction is guided by international experts	The data collection methodology is not discussed	The data collection methodology is not discussed	The data collection methodology is not discussed	The data collection methodology is not discussed	The data collection methodology is not discussed	The questionnaire construction is consistent with previous versions of the YRBS	Cannot be determined as the dataset is not published	The questions covered follow the 8 studies by subject area experts and shares many of Statistics South Africa's conventions	The questions covered follow international TIMSS norms	The metadata does not state whether international standards are adopted
			Does the metadata describe changes to the methodology and/or rules, definitio	The metadata discusses changes to the structure of questions and available options to respondents of the survey	The metadata discusses changes to the structure of questions and available options to respondents of the survey	The metadata discusses changes to the structure of questions and available options to respondents of the survey	The data collection methodology is not discussed	The data collection methodology is not discussed	The data collection methodology is not discussed	The data collection methodology is not discussed	The data collection methodology is not discussed	There were no changes made to interpretations of terms	Definitions are not provided in the documents	The metadata discusses changes to the structure of questions and available options to respondents of the survey	The metadata discusses changes to the structure of questions and available options to respondents of the survey	The metadata does not highlight changes to the structure of questions and available options to respondents of the survey

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			ns that are used?	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟩	🟥
	Is the data structurally consistent?	Are data values consistently reported to the public?	Are data values over time are reported consistently?	The data matches the results documentation	The data matches the results documentation	The data matches the results documentation	Data is consistently reported over the period of available data	Data is consistently reported over the period of available data	Data is consistently reported over the period of available data	Detailed statistics and metadata is not released to make a determination	Detailed statistics and metadata is not released to make a determination	The data matches the results documentation	Cannot be determined as the dataset is not published	The data matches the results documentation	The data matches the results documentation	A findings report is made available referring consistently to the fields included in the codebook
			Does the data provider provide reasons for changes in data values?	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms	If there were changes, it is not documented and therefore one is unaware if there were issues	If there were changes, it is not documented and therefore one is unaware if there were issues	If there were changes, it is not documented and therefore one is unaware if there were issues	If there were changes, it is not documented and therefore one is unaware if there were issues	If there were changes, it is not documented and therefore one is unaware if there were issues	There were no changes made to interpretations of terms	Cannot be determined as the dataset is not published	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms	There were no changes made to interpretations of terms
Is the data current enough and released in a timely manner for the data user?	Is the dataset punctually released?	Is the dataset released in a punctual manner?	Is the average time between the end of the data collection and the data release within recommended time-frames?	Data is released timeously	Not applicable - Detailed data is not ready for publication	Data is released timeously	Crime statistics are released over a year after collation and the release is not scheduled	It is not clear when the master list is updated each year.	Education statistics are released over a year late in findings reports and the release is not scheduled	Education statistics are released over a year late in findings reports and the release is not scheduled	Education statistics are released over a year late in findings reports and the release is not scheduled	Detailed data is not published; it was provided on request only	Survey results released 4 years after collection	Data is released timeously	South African TIMSS data is embargoed for 2 years before release due to HSRC data policy	SASAS data is embargoed for 2 years before release due to HSRC data policy
			Does the organisation release a report on the time frame for data releases?	The census is carried out every 10 ten years, but a schedule of data outputs is provided	Not applicable - The date for the next CS has not been scheduled as yet - is 10 years away	The GHS is carried out every year and data is schedule for release	A time frame for reporting is not provided	A time frame for reporting is not provided	A time frame for reporting is not provided	A time frame for reporting is not provided	A time frame for reporting is not provided	The survey is conducted irregularly	There is no schedule of surveys	The survey is carried out every 2 years	The survey is carried out every 4 years	The survey is carried out every year but with no schedule
Is the data accessible to the data user?	Are data access controls too stringent?	Are the applicable data users granted an appropriate level of data access?	Are the access rights clearly and uniformly applied	Data is freely available for download off the website. Small area data can be purchased	Not applicable - Detailed data is not ready for publication	Data is freely available for download off the website. ASCII datasets are	Freely available - no limitations, detailed information is unavailable	Freely available - no limitations	Data access to individual surveys is strongly guarded. Access by request only	Data access to individual surveys is strongly guarded. By access request only	Data access to individual surveys is strongly guarded. By access request only	Data access is limited and not discussed on the MRC website	Datasets are not published. No indication is provided on the website	There are 3 access level. Public, internal and secure. Secure and Internal is	The HSRC controls data access. Registration is required on the HSRC website	The HSRC controls data access. Registration is required on the HSRC website

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			ensuring that the correct individuals are granted access?	from Statistics South Africa		available on request							how to access the data	only accessed within SALDRU and is for operational purposes	before one can download the data	before one can download the data
			Can users easily find directions on how to access the required information?	Information is available on the Statistics South Africa website and a handbook is provided explaining how to access data	Not applicable - Detailed data is not ready for publication	Information is available on the Statistics South Africa website	No supporting documentation is provided	No supporting documentation is provided	No supporting documentation is provided	No supporting documentation is provided	No supporting documentation is provided	Data access is limited and not discussed on the MRC website	Data is only available in the statistics reports	Guidance for requesting data is clearly provided and the steps to do so are simple to follow	Guidance for requesting data is clearly provided and the steps to do so are simple to follow	Guidance for requesting data is clearly provided and the steps to do so are simple to follow
			Does the data provider respond to data requests in line with international standards on responsiveness	Data requests are met in a timely manner	Data requests are met in a timely manner	Data requests are met in a timely manner	Data is available for download; no additional user support is provided for more detailed information	Data is available for download; no additional user support is provided for more detailed information	Data is available for download; no additional user support is provided for more detailed information	Data is available for download; no additional user support is provided for more detailed information	Data is not available for download; data provider was not found to be responsive to requests	Data access is limited and not discussed on the MRC website	Not applicable - no channel for contacting MRC resources is provided to request such data	Data requests are met in a timely manner	Data access requests are met in a timely manner	Data access requests are met in a timely manner
	Is a suitable granularity of data accessible to the public?	Are access controls applied to the correct fields and categories?	Are access controls applied to the correct selection of fields and or categories?	All data apart from individual's personal details are made available, following national legislation	Not applicable - Detailed data is not ready for publication	All data apart from individual's personal details and geography details lower than Metro name are made available	Granular data is made available to police station level	Granular data is made available to school level	Granular data is not published for no particular reason	Granular data is not published for no particular reason	Granular data is not published for no particular reason	Data access is limited and not discussed on the MRC website	Data is only available in the statistics reports	All data apart from individual's personal details and location are made available	All data apart from individual's personal details and location are made available	All data apart from individual's personal details and location are made available
			Adequate levels of anonymity need to be applied whilst ensuring that the user's	Data has been anonymised	Not applicable - Detailed data is not ready for publication	Data has been anonymised	Data is aggregated to municipal level ensuring anonymity	Not applicable	Anonymity of data is wrongly applied	Anonymity of data is wrongly applied	Anonymity of data is wrongly applied	Respondents names are excluded	Data is aggregated in the survey report, thus individual particulars are excluded	Data has been anonymised	Data has been anonymised	Data has been anonymised

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			needs for data are considered	🟩	⬛	🟩	🟩	⬛	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩
			Does the metadata describe the levels of detail that are accessible?	The results documentation discusses the detail available within the Census, supporting codes and descriptions are available for download	Not applicable - Detailed data is not ready for publication	The results documentation discusses the detail available within the GHS supporting codes and descriptions are available for download	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	The metadata does not describe the geographic details or the file structures	The metadata does not describe the geographic details or the file structures	The results documentation discusses the detail available within the Census, supporting codes and descriptions are available for download	The results documentation discusses the detail available within the survey	The level of detail of data access available is documented within the metadata
	Is the data and metadata provided in a meaningful and useful manner?	Is data and metadata provided in a useful and meaningful manner?	Is the data dissemination strategy shared with the public?	The metadata details how to access the data	Not applicable - Detailed data is not ready for publication	There is no specific data dissemination strategy that is published	There is no published data dissemination strategy, but the data is released publicly	There is no published data dissemination strategy, but the data is released publicly	The data dissemination strategy is not shared with the public	The data dissemination strategy is not shared with the public	The data dissemination strategy is not shared with the public	There is not published data dissemination strategy	There is no published data dissemination strategy	The metadata details how to access the data	The metadata details how to access the data	The metadata details how to access the data
			Is the data release scheduled?	The next census is schedule for 10 years. A mini census (Community Survey) was released in 2016	The detailed data is due in late 2016	The GHS is annual is schedule for release	The next release is not dated	The next release is not dated	The next release is not dated	The next release is not dated	The next release is not dated	The next release is not dated	The next release is not dated	Data is set for collection every 2 years. The exact date of the next release is not discussed	Data is set for collection every 4 years. The exact date of the next release is not discussed	The survey is carried out every year but with no schedule
			Does the metadata provide guidance on how to access and retrieve data using the available mediums?	The metadata provides steps on how to access and retrieve data	Not applicable - Detailed data is not ready for publication	The website discusses how to access data but documents do not detail how one accesses such data in detail	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	No metadata is provided	Tables are provided without guidance on how to use the various mediums available	Tables are provided without guidance on how to use the various mediums available	The metadata provides steps on how to access and retrieve data	The metadata provides steps on how to access and retrieve data	The metadata provides steps on how to access and retrieve data
			Is the data dissemination strategy frequentl	The documentation follows the current accessibility of data	Not applicable - Detailed data is not ready for publication	There is no specific data dissemination strategy that is published	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	The documentation follows the current accessibility of data	The documentation and data release follows the HSRC data	The documentation and data release follows the HSRC data embargo policy

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			y updated?												embargo policy	
			Does the data dissemination strategy take heed of the needs of the data users?	The data dissemination strategy is in line with collected data requirements of the public	Not applicable - Detailed data is not ready for publication	There is no specific data dissemination strategy that is published	There is no data dissemination strategy that is published	There is no data dissemination strategy that is published	There is no data dissemination strategy that is published	There is no data dissemination strategy that is published	There is no data dissemination strategy that is published	Database tools could be adopted to improve usability when interacting the data. Data is not made available	The needs of the public do not seem to be factored into consideration	The data dissemination strategy is in line with collected data requirements of the public	Data dissemination follows HSRC data policy not user requirements	Data dissemination follows HSRC data policy not user requirements
	Is the channel of data delivery appropriate for the data user?	Is the channel of data delivery suitable for the needs of the user?	Are hardware and software requirements clearly communicated to the data user?	Software and hardware requirements are discussed	Not applicable - Detailed data is not ready for publication	Software and hardware requirements are discussed on the website	Software and hardware requirements are not explicitly discussed	Software and hardware requirements are not explicitly discussed	Software and hardware requirements are not explicitly discussed	Software and hardware requirements are not explicitly discussed	Software and hardware requirements are not explicitly discussed	Software and hardware requirements are not discussed	Software and hardware requirements are not discussed	Software and hardware requirements are discussed	Software and hardware requirements are discussed	Software and hardware requirements are discussed
			Have the user's preferred means of accessing data been taken into consideration when developing the data access channel?	The data dissemination strategy is in line with collected data requirements of the public	Not applicable - Detailed data is not ready for publication	The users' means of access follows the requirements stated in the Census	The channel could be improved with database tools, but Excel is a common preference for distributing data	The channel could be improved with database tools, but Excel is a common preference for distributing data	Data users are not strongly considered	Data users are not strongly considered	Data users are not strongly considered	The channel could be improved	The needs of the public do not seem to be factored into consideration	The data dissemination strategy is in line with collected data requirements of the public	User's preferences are not determined in terms of data sharing	User's preferences are not determined in terms of data sharing
			Are the disseminated datasets accessible in multiple data formats?	Data is also provided in multiple formats, apart from Super Cross and Nesstar tools	Not applicable - Detailed data is not ready for publication	Data is also provided in multiple formats, apart from Super Cross and Nesstar tools	Provided only in Excel	Provided only in Excel	Provided only via Statistical PDF reports	Provided only via Statistical PDF reports	Provided only via Statistical PDF reports	Provided only in SPSS	Provided only in the survey report	Data is also provided in Excel, Microdata and Stata	Data is also ASCII, SAS, SPSS and STATA	Data is also ASCII, CSV, SAS, SPSS and STATA

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			Where database tools are provided, is the mode of access understandable and simple to operate?	The database tools are discussed within metadata	Not applicable - Detailed data is not ready for publication	The database tools are not discussed within metadata	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable - SALDRU does not provide online database tools	Not applicable - HSRC does not provide online database tools	Not applicable - HSRC does not provide online database tools
			Are user manuals provided to aid data access to database tools?	User manuals provided to aid data access to database tools	Not applicable - Detailed data is not ready for publication	User manuals are not provided to access to the database tools	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable - SALDRU does not provide online database tools	Not applicable - HSRC does not provide online database tools	Not applicable - HSRC does not provide online database tools
			Where external staff resources need to be contacted to request dataset, are such staff reachable?	Statistics South Africa personnel are responsive to data requests	Not applicable - Detailed data is not ready for publication	Statistics South Africa personnel are responsive to data requests	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	Not applicable	External staff are responsive in providing data access	External staff are responsive in providing data access	External staff are responsive in providing data access
Does the data have integrity when consumed by the data user?	Are data quality constraints suitable for the needs of the data user?	Do the data quality constraints applied within the data system sufficient to manage data integrity?	Do the data quality constraints follow international best practices?	Attention is given to ensure that fieldworkers understand definitions and that data quality constraints are introduced to the data collection strategy	Attention is given to ensure that fieldworkers understand definitions and that data quality constraints are introduced to the data collection strategy	Attention is given to ensure that fieldworkers understand definitions and that data quality constraints are introduced to the data collection strategy	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	Data input constraints are applied to the electronic form capture used in the survey	Data input constraints are not discussed within the documentation	Attention is given to ensure that fieldworkers understand definitions and that data quality constraints are introduced to the data collection strategy	Attention is given to ensure that fieldworkers understand definitions and that data quality constraints are introduced to the data collection strategy	Data constraints are not discussed within the metadata
			Are the data rules suitably captured within	Only data values as discussed within the metadata are provided	Only data values as discussed within the metadata are provided in	Only data values as discussed within the metadata are provided	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	Only data values as discussed within the metadata are provided	Data input constraints are not discussed within the	Only data values as discussed within the metadata are provided	Only data values as discussed within the metadata are provided	Data constraints are not discussed within the metadata

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			the data quality constraints?		Statistical Release document								documentation			
			Is the process to address data quality breaches discussed within the metadata?	The process for resolving data quality breaches are detailed within the Methodology	The process for resolving data quality breaches are detailed within the Methodology	The process for resolving data quality breaches are detailed within the Methodology	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	Data quality constraints are not discussed	The process for resolving data quality breaches are not detailed within the Methodology	The process for resolving data quality breaches are not detailed within the Methodology	The process for resolving data quality breaches are detailed within the Methodology	The process for resolving data quality breaches are detailed within the Methodology	The process to address data quality breaches are not provided within the metadata
	Does the data provider introduce the most appropriate statistical and methodological approaches?	Are the data transformation techniques suitable?	Are the applied data transformation techniques applied within the bounds of international best practices?	The data aggregations are accepted by the panel of experts	The data aggregations are accepted by the panel of experts	The data aggregations are accepted by the panel of experts	There is no metadata to assess the data transformation techniques applied	There is no metadata to assess the data transformation techniques applied	There is no metadata to assess the data transformation techniques applied	There is no metadata to assess the data transformation techniques applied	There is no metadata to assess the data transformation techniques applied	Not applicable - The data provided is not aggregated	The data transformation techniques are not discussed	All data releaseed is vetted by a Commerce Faculty Ethics Committee	All data transformations follow TIMSS standards	Data transformations are not discussed within the metadata
			Are the data transformation methodologies detailed within the supporting metadata?	Aggregation of data matches the metadata	Not applicable - Detailed data is not ready for publication	Aggregation of data matches the metadata	There is no metadata to assess the data transformation techniques applied	There is no metadata to assess the data transformation techniques applied	There is no metadata to assess the data transformation techniques applied	There is no metadata to assess the data transformation techniques applied	There is no metadata to assess the data transformation techniques applied	Not applicable - The data provided is not aggregated	The data transformation techniques are not discussed	Aggregation of data matches the metadata	Aggregation of data values matches the data reports that are published	Data transformations are not discussed within the metadata
Is the data traceable?	Does the data provider track the data source?	Does the organisation track data transformations?	Does the organisation document the data sources and transformations	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	There is no metadata to assess the quality of the source systems that are used	There is no metadata to assess the quality of the source systems that are used	There is no metadata to assess the quality of the source systems that are used	There is no metadata to assess the quality of the source systems that are used	There is no metadata to assess the quality of the source systems that are used	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required

Dimension	Element	Indicator	Focal Issue	Census of South Africa 2011	Community Survey 2016	General Household Survey 2015	SA Crime Statistics	School Master List	Annual School Survey 2014	NEIMS 2016	Annual National Assessment 2014	YRBS 2011	Demographic and Health Survey 2003	National Income Dynamics Study Wave 1,2,3,4	Trends in Mathematics and Science Survey 2011	South African Social Attitudes Survey 2012
			which are used within the metadata ?													
			Has supporting information been vetted using the public data quality assessment framework?	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	There is no metadata to assess the quality of the source systems that are used	There is no metadata to assess the quality of the source systems that are used	There is no metadata to assess the quality of the source systems that are used	There is no metadata to assess the quality of the source systems that are used	There is no metadata to assess the quality of the source systems that are used	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required	Not applicable - no source systems required