RESEP
Research on Socio-Economic Policy

The quality and importance of SA-SAMS data as unit-level data: a technical overview

January 2023

Chris van Wyk

Stellenbosch Working Paper
Series No. WP01/2017
Publication Date: 2016 Keywords:
Labour, Education, Health
JEL Codes: B21, G10

Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Department of
Economics Matieland
7602

THE QUALITY AND IMPORTANCE OF SA-SAMS DATA AS UNIT-LEVEL DATA: A TECHNICAL OVERVIEW

1. MANAGEMENT OF THE DATA

The data used in this report is at the learner unit record level and is based on data predominantly from the DDD, SA-SAMS, LURITS and NSC obtained from the MSDF, the EMIS section of three provinces (GT, EC, and LP) and DBE as indicated in **Table 1** below. Learner unit record data refers to the data collected for each learner through the South African School Administration and Management System (SA-SAMS).

Table 1: Datasets from the relevant organisations

Datasets	***Purpose***	***Comment***
DDD (3 Provinces)	*Cohort Analysis (Before and After Covid-19)*	*Linked and used learner unit level data from schools that submitted every year from 2016 to 2022 using a unique identifier (New advanced algorithm)*
DDD (3 Provinces)	*Enrolment Patterns (Repetition, Dropout)- (Before and After Covid-19)*	*Linked and used learner unit level data from schools that submitted every year from 2016 to 2022 using a unique identifier*
LURITS (SA)	*Enrolment Patterns (Repetition, Dropout)*	*Linked and used learner unit level data from schools that submitted every year from 2018 to 2021 using a unique identifier*
DDD (3 Provinces)	*School Based Assessments*	*Used learner unit level subject data from 2016 to 2021 (The linking of SBAs performance over time using a common field makes it possible to start investigating SBAs across schools, and how this influences learner outcomes and learner flows)*
SA-SAMS for EC, LP and GT	*Age Distribution and Overage*	*Used the learner unit level data from the learner table in SA-SAMS database of the current year*
SA-SAMS for EC, LP and GT	*Teacher Details*	*Used the individual teacher data in the Educator table in SA-SAMS database of the current year*
SA-SAMS for EC, LP and GT	*Subjects Taught by Teachers*	*Used the individual teacher data: Linked the current educator table with the Educator Subject Taught table in SA-SAMS database*
SA-SAMS GT	*Feeder Schools*	*Used learner unit level data: Linked the current Gr 8 learners with the Gr7 learners of the previous year from the Learner tables in SA-SAMS database.*
NSC data from DDD	*SBA data predicts NSC outcomes*	*Matching individual matric examination data to the DDD SBA data through a unique identifier (anonymised SA ID)*
NSC from DBE	*NSC Performance over time*	*Analysis of NSC data from 2008 to 2021*
Master list of Schools	*Integration of data sets and providing relevant details of schools*	*Uniquely identify each school in the country through a school identifier, generally called the "EMIS number".*

The datasets generated by EMIS in most countries in the world are often some of the most under-utilised data sources. The objective of an education management information system (EMIS) is not only to collect, store and process information but also to help in education policymaking, by providing relevant and accessible information for research projects such as required by this project. The aim of this section is not only to focus on the results of the analysis but to emphasize the importance and quality of unit record data and the power of longitudinal data.

There are various major challenges that come into the way while dealing with big data. Managing and analysing large amounts of data remains a big challenge. The *volume* of this learner unit record data is really a big deal in managing a dataset of this size. A further challenge is the speed to process and manage the data. Big data also increases the difficulty of data integration. Data integration in this context is to link the data from different years to follow individual learners over time. Therefore, analysing and processing big data sets remains a challenge to get the data in a format and structure that is manageable. It was possible to format and analyse the data because of the availability of a detailed entity relationship diagram (ERD)[1] which we obtained from the database developers.

An Entity Relationship Diagram (ERD) is a type of flowchart that gives a snapshot of how the entities (schools, learners, teachers, subjects, etc.) relate to each other. It is the blueprint that gives a visual representation of the relationships between the different sets of data (entities). An entity-relationship diagram is essential for modelling the data stored in a database. It is the basic design upon which a database is built and shows how entities relate to other entities and should be easily available and accessible to the data users. The ERD's in **Figures 1 and 2** show how to link a learner to their school and to their subjects in SA-SAMS and DDD database systems.

[1] An **entity relationship diagram** (ERD) is a graphical representation of an information system that depicts the **relationships** among the tables within the database system.

Figure1: ERD for SA-SAMS

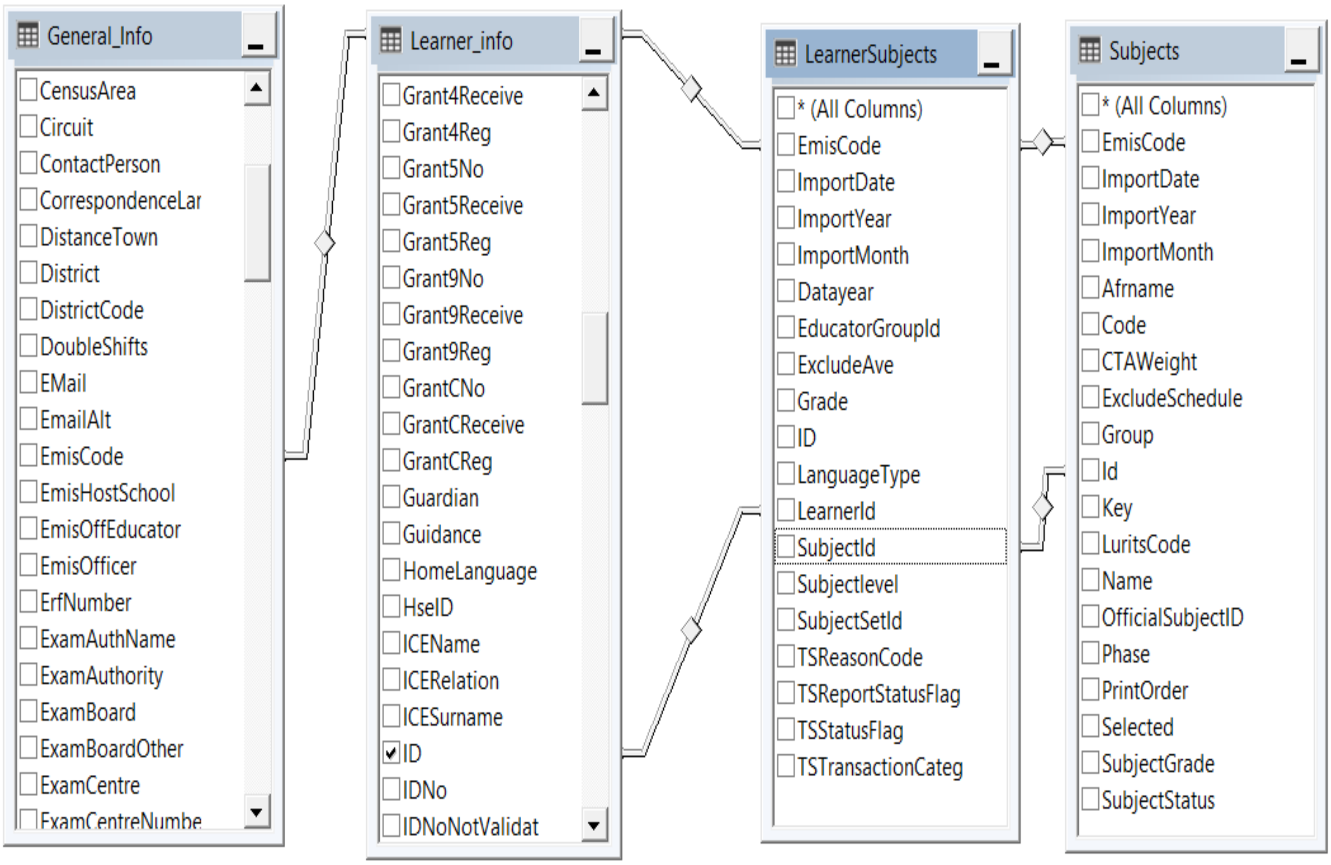
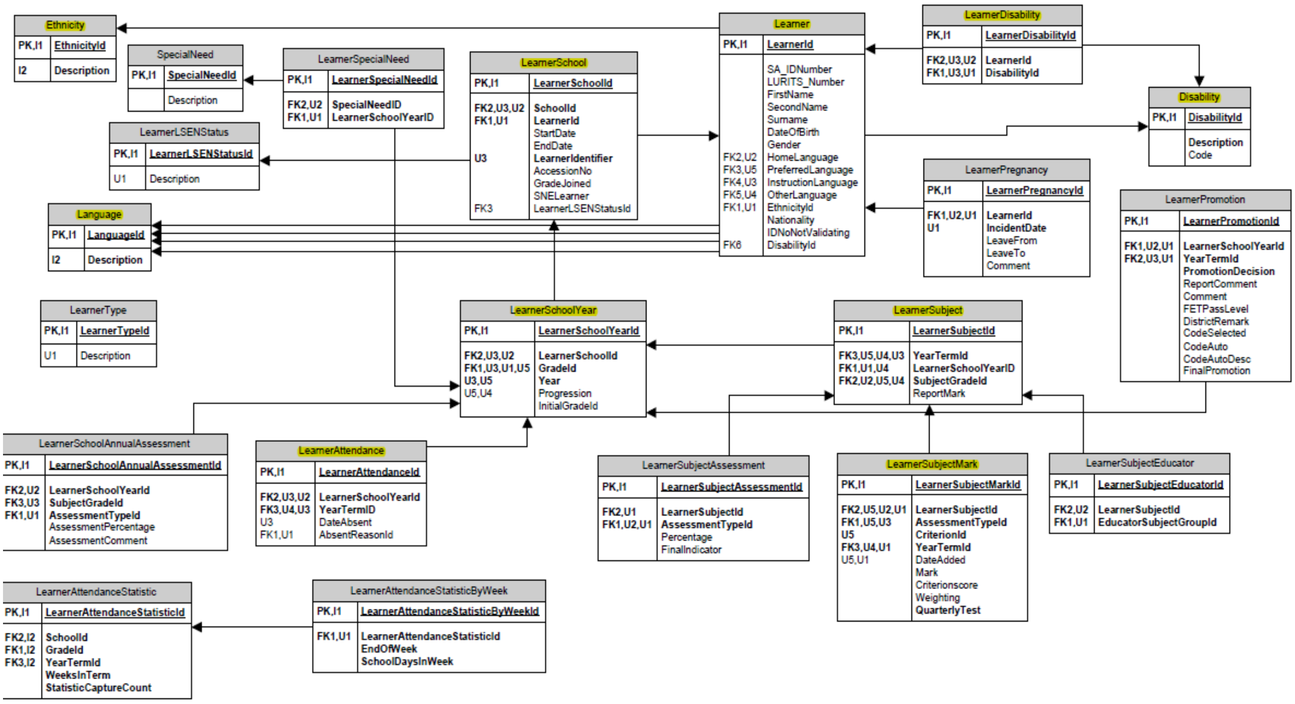


Figure 2: ERD for DDD

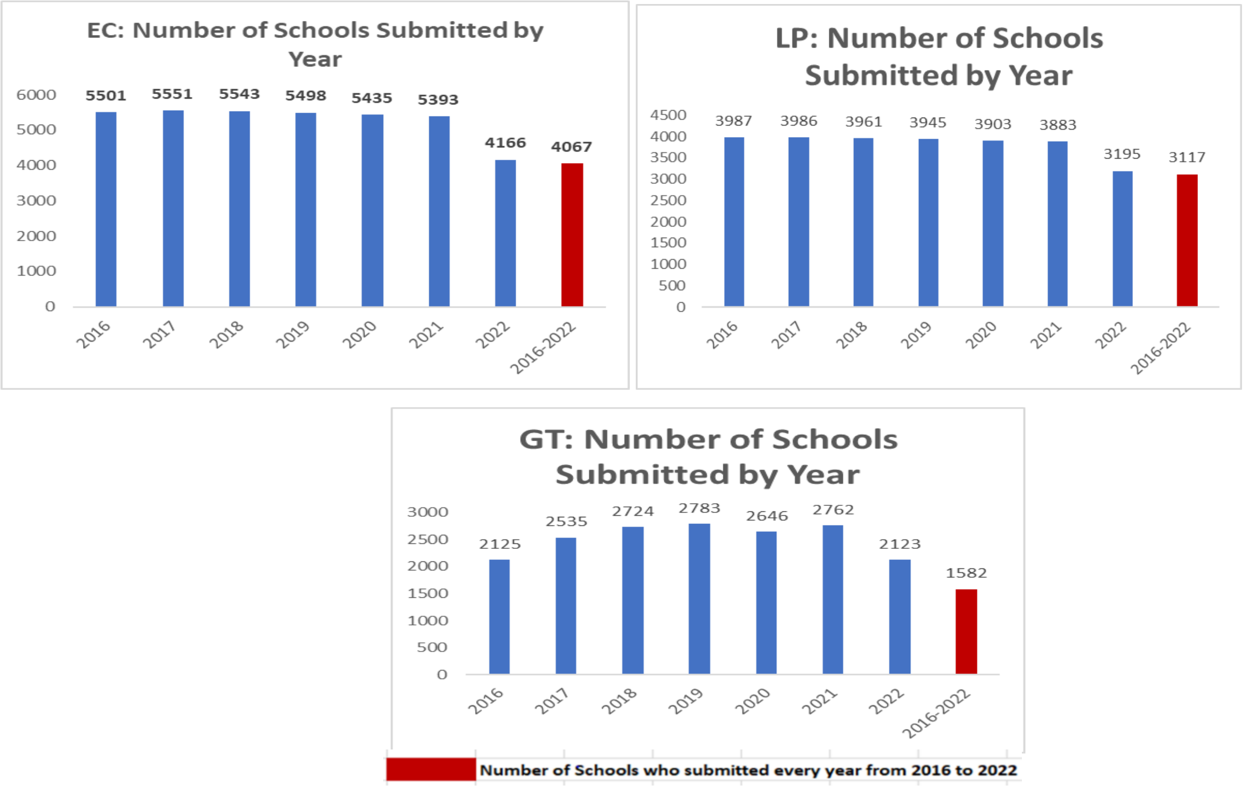


2. DATA QUALITY ISSUES

The quality of data is determined by factors such as accuracy, completeness, reliability, relevance, and timeliness. The focus of this assessment is not on the results but rather on the methodology that could be applied to improve the quality of the data in SA-SAMS. Here the level of data quality is measured against three dimensions: Completeness, Accuracy and Consistency. How good is the quality of the data generated by SA-SAMS? The data received have some quality-related issues:

2.1. Data Inconsistencies:

Figure 3: Number of schools that submitted data by Year in three Provinces (EC, GT, LP)



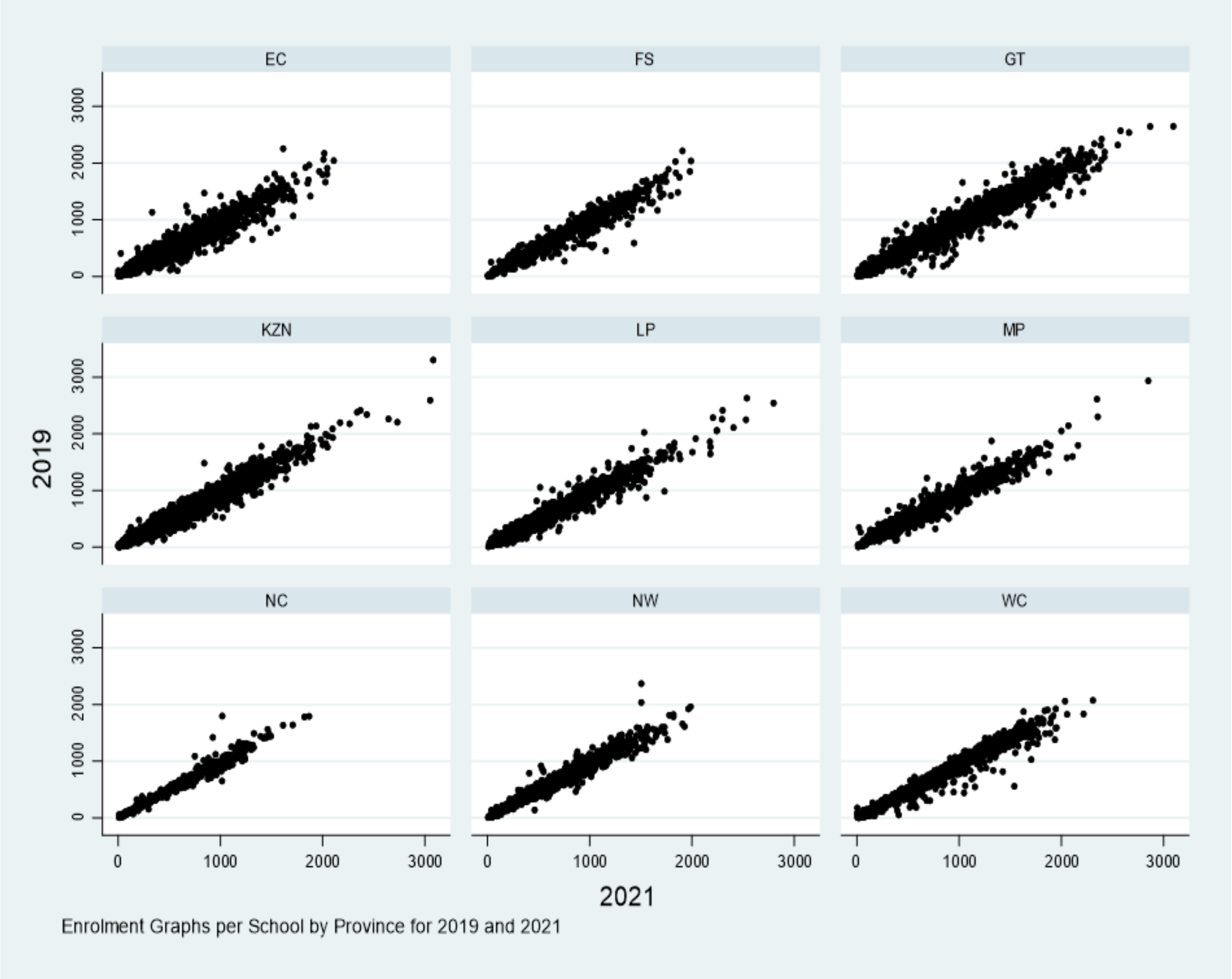
Data Source: DDD

Data consistency is considered as one of the important dimensions of data quality. **Figure 3** shows the number of schools that submitted data from 2016 to 2022 vs the total of schools that submitted every year for that period (indicated in red) in 3 provinces. **Figure 3** shows that there is a vast difference between the number of schools that submitted data annually versus the number of schools that submitted every year since 2016. It is important to use the same schools over time in the cohort analysis otherwise the dropout could be over-reported. To remedy this problem, we used the same schools that submitted every year from 2016 to 2022 in our analysis. The result is that we have fewer schools in the analysis. Following this approach, a selection bias is introduced by the selection of the same schools that submitted every year. Because we do not have information for all schools, we cannot tell whether someone really dropped out [2]or moved to a school that was not recorded.

[2] Dropped out in this report referred to the unaccounted learners who were not in the GDE system anymore.

However, for the schools that submitted the quality of the data is of good standard (especially in 2018 to 2021). **Figure 4** shows the comparison of enrolment in LURITS between 2019 and 2021 for all provinces in the country. The aggregated LURITS data (at a school level) seemed to be of better quality than unit-level (learner-level) data as indicated by Figure 4.

Figure 4: Scatterplots for Enrolment per school by province for 2019 and 2021



Response rate of enrolment

Table 2: Response rate of schools who submitted learner data in 2021

	GT	LP	EC
DDD2021	2762	3883	5393
Masterlist2021	3008	3858	5752
Response Rate	92%	101%	94%

Figure 5: Number of schools by Province and Category Type

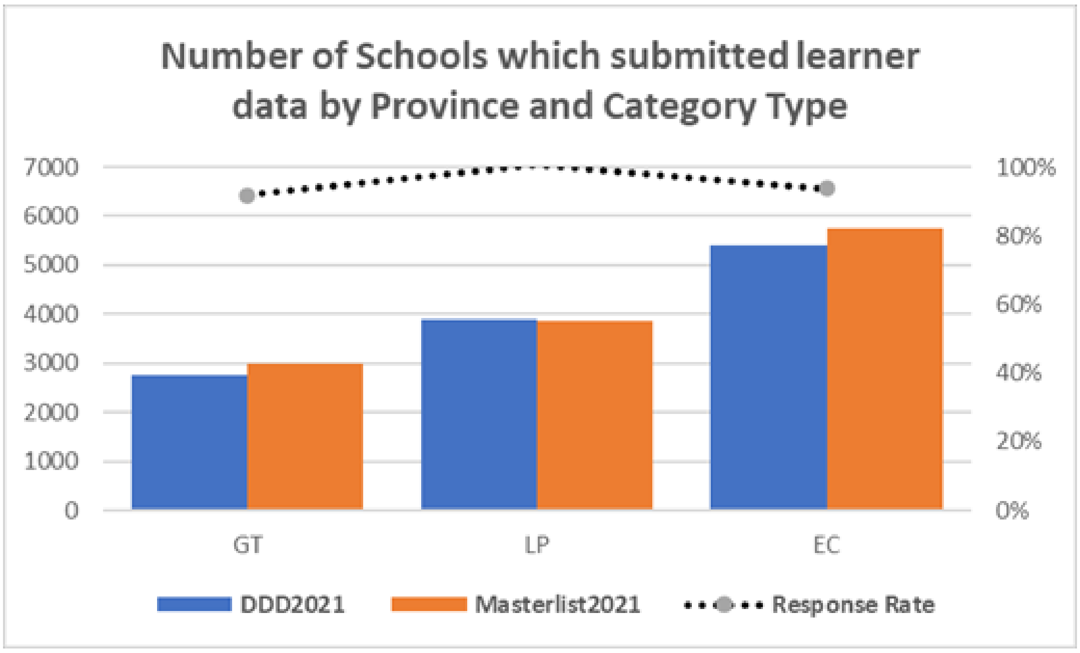
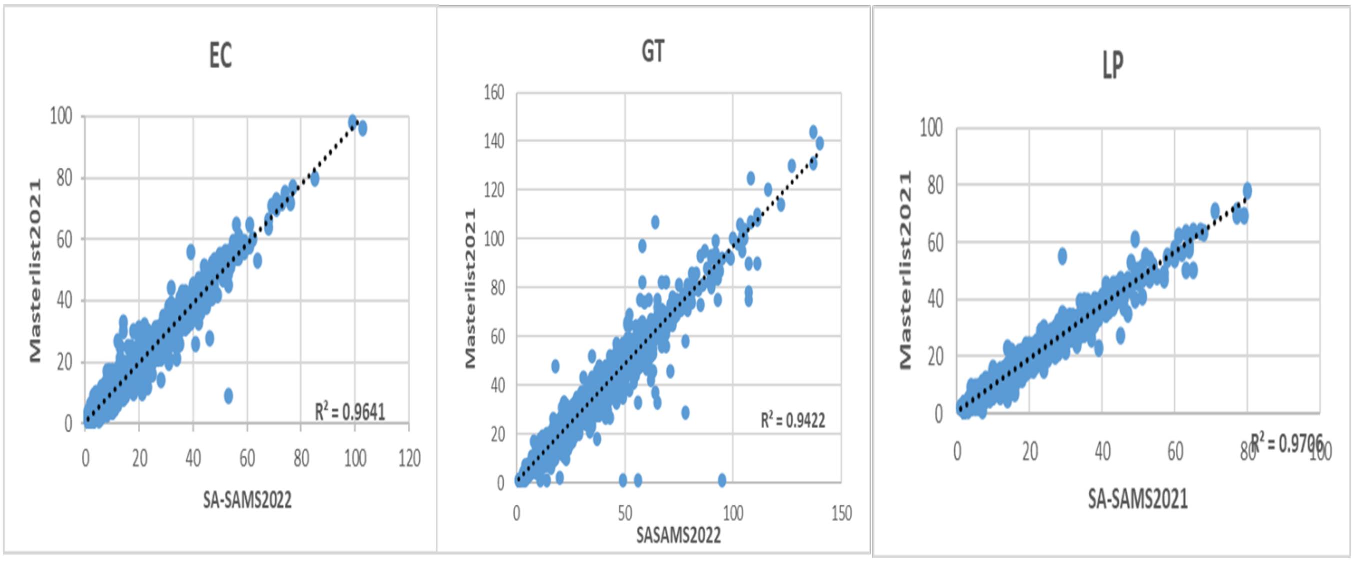


Table 2 and **Figure 5** indicate the number of schools that submitted learner data in DDD database and the number of schools in the Master list of schools in 2021. The last row in the table shows the response rate of the schools in comparison with the 2021 master list of schools who submitted learner data. The response rate is one of the factors that could potentially influence data quality, although a high response rate is not a guarantee of high-quality data. The response rate in all three provinces is relatively high. *Completeness*, as earlier indicated, is a measure of data quality. When we talk about data completeness, the most common situation we encountered is empty cells in a data table. For example, if the learner date of birth is missing, data can be considered incomplete. Completeness measures if the data is sufficient to deliver meaningful inferences and decisions. Date of birth is such an important data element. With the date of birth, the learner's age can be calculated. Age is more than just a number; age is important to determine whether the learner is overage and overage learners give an indication of repetition in the system. *Accuracy* is another dimension of data quality. It is of no use the learner's date of birth is completed, but it is wrong. Inaccurate birth details give inaccurate age distribution.

Response rate of teachers

Most impressive was the response rate of the table for individual teachers. According to the teacher table in SA-SAMS, schools submitted detail of their individual teachers. **Figure 6** shows the relationship between the number of teachers per school in SA-SAMS and the number of teachers per school in the Master list. As indicated in the Figure there is a very strong positive relationship between the schools in the master list and the table in SA-SAMS in all three provinces.

Figure 6: Scatterplots for EC, GT and LP between the number of teachers in SA-SAMS per school and the number of teachers per school in the Master list



An important and unique dataset in SA-SAMS is the table that indicates the subjects and grades taught by teachers. This is an important curriculum related module in SA-SAMS (DBE,2022). This includes the recording and reporting on the academic progress of learners. In this module the learners and teachers are allocated subjects and placed into classes (available from: https://sasams.co.za/). According to the SA-SAMS user manual (Available from: https://sasams.co.za/repo/modules/12.pdf) "The Setup Subject and Subject Choices menu is used to set up the curriculum framework of the school. It allows the user to manage the subjects offered by the school, assigns subjects to the learners, and creates subject groups per educator". However, when we formatted and analysed this data it seemed that it is not accurate and complete. A high percentage of teachers in GT and LP did not complete this table. Furthermore, result of the class variable per school also seemed to make no sense.

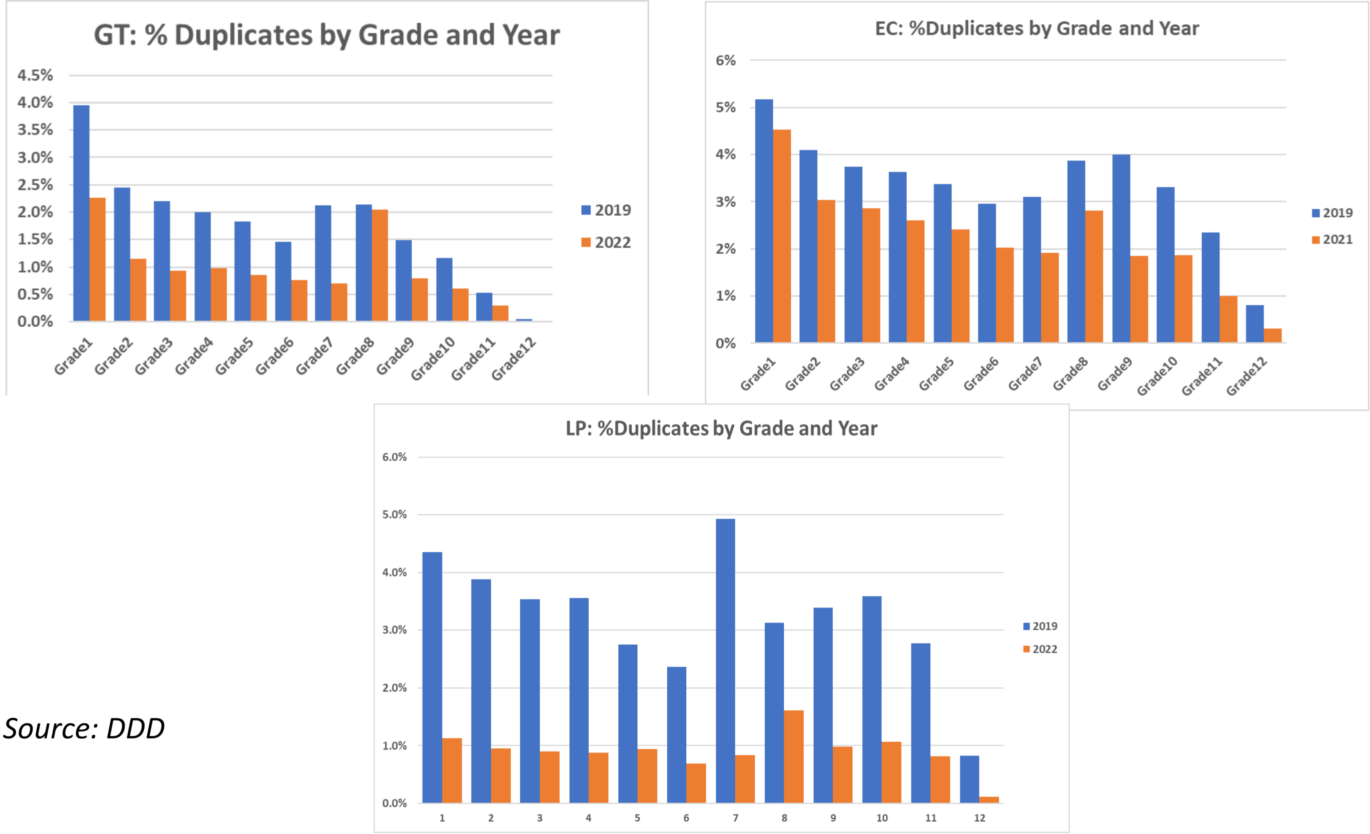
As a guiding principle and as a recommendation we recommend that data verification at district, provincial and national levels, is not only focused on response rate but that it also focuses on data completeness and accuracy.

2.2. Duplicate Learners:

Duplicate learners in the DDD learner data

We identified, based on the learner's unique identifier, several learners in the DDD data who appeared more than once in the same year. The reason could be that learners are probably moving between schools. Such examples don't really affect the analysis because we simply will take the last school attended. However, the duplicates where we have the same learner in more than one grade in the same year is somewhat problematic and it appears to be a data error. The lack of consistent unique identifiers is also a direct result that duplicate learners appear in the dataset.

Figure 7: Duplicate learners by Grade and Year in 3 Provinces (GT, LP, EC)



Source: DDD

Figure 7 shows the duplicate learners in three provinces (GT, EC and LP) for 2019 and 2021. These figures indicate that there was an improvement in the number of duplicates in all three provinces between 2019 and 2021. This could be attributed to the fact that DDD introduced an enhanced algorithm to create a unique identifier. The purpose of DDD Learner Matching Algorithm was to match learner data submitted by schools (from SA-SAMS) to existing learners within the DDD system. This allows for the tracking of learners within schools, within the system and learner movement between schools.

Duplicate identifiers in the DDD subject data

With the inclusion of learners' anonymized South African ID numbers in the DDD dataset, it is possible to investigate if each learner is allocated to a single unique identifier, or if duplicate identifiers exist for a single learner. **Table 3** below shows what percentage of the total observations in the DDD dataset are missing a South African ID number and what percentage of South African ID numbers are assigned to two or more unique learner identifiers.

Table 3: Percentage Duplicate and Missing Learners by Province

	GT	EC	LP
2+ Learner IDs	5%	6%	11%
Missing	22%	8%	5%

Source: DDD subject data

While some of the observations attached to a single South African ID number appear to be errors, the majority look as though they represent a single learner which is represented by two “unique” observations in the data. This can be seen by comparing the home language, year of birth and gender of observations with the same South African ID number, and then by seeing if the missing grades for each observation align with the non-missing grades for the accompanying observation(s).

Where the SA ID number was missing for an observation, it was impossible to determine the extent of duplicate identifiers. This may be the reason that Gauteng has the lowest percentage of 2+ Learner IDs, as they have the lowest proportion of observations with a SA ID number. However, it is impossible to know with the current data if this is the case. The above highlights the importance of learners’ SA ID numbers being captured (and captured correctly), as this is likely the most reliable learner identifier and if verified against a database, it will less likely than learner information to be captured incorrectly, which is likely the cause of a single learner being allocated multiple unique learner identifiers.

2.3. Missing Data

A key element of the project is to determine the correlation between performance in internal school assessments and external matric examinations. However, this was not possible in previous years because NSC results were not included in the dataset for any province. Furthermore, our analysis thus far has showed that SBA provides considerable information that predict later outcomes. Now that the individual matric examination data is included and matched to the DDD SBA data would improve our ability to analyse the quality and consistency of the SBA data. Learners in some schools and provinces (depending on the data collection) can now potentially be followed from Grade 8 to Grade 12 (2016-2021), with Grade 12 SBA and matric results forming part of the analysis. Further analysis of how SBA performance predicts matric performance can help to assist in measuring the quality and leniency of SBA. This is an exciting prospect, given how little research has been done thus far on the quality of SBA data.

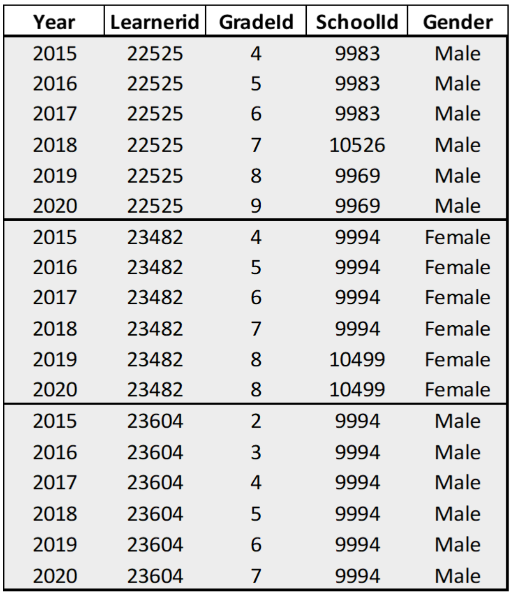
3. UNIQUE IDENTIFIERS:

A unique identifier is a single, non-duplicated number that is assigned to, and remains with, a learner throughout his or her education career irrespective of whether the learner changes schools (CPSI Ltd., 2010). It is of utmost importance that a unique identification code must be assigned to every learner. It is important that this identifier is consistent and accurate over time. The National Idenfication Number is the ideal number to be used for such a purpose. The allocation of a unique identification number in SA-SAMS to all learners that is consistent for all years is probably the single and most important limitation.

In creating a longitudinal data system, it is necessary to link the different datasets that have been collected for individual learners or individual schools for each year by using a common field across these datasets. The principle in SA-SAMS is that a learner is only unique within a school. This works well in terms of the management of the database at a technical level or if a learner never changes schools. However, it is problematic for longitudinal data coverage when one wants to gather data for the same learner from year to year. Longitudinal analysis is particularly useful when used for cohort analysis. Longitudinal data allow us to study and understand patterns in education over time and, crucially, across grades. This can be particularly useful for the study of education flows, as

longitudinal data can shed light on issues such as repetition and dropout in education. **Table 4** shows the format of longitudinal data (data of the same learners from year to year). This longitudinal data format enables us to determine grade progression and repetition; school switching (feeder schools) and dropout as outlined in the following paragraphs.

Table 4: Data format of Longitudinal data (Data on multiple units at multiple points in time)



Year	Learnerid	GradeId	SchoolId	Gender
2015	22525	4	9983	Male
2016	22525	5	9983	Male
2017	22525	6	9983	Male
2018	22525	7	10526	Male
2019	22525	8	9969	Male
2020	22525	9	9969	Male
2015	23482	4	9994	Female
2016	23482	5	9994	Female
2017	23482	6	9994	Female
2018	23482	7	9994	Female
2019	23482	8	10499	Female
2020	23482	8	10499	Female
2015	23604	2	9994	Male
2016	23604	3	9994	Male
2017	23604	4	9994	Male
2018	23604	5	9994	Male
2019	23604	6	9994	Male
2020	23604	7	9994	Male

4. THE POWER OF LONGITUDINAL DATA

Aggregated Data

Aggregated data refers to data collected at the school level (the school census approach).

The aggregated data collected through the Annual School Survey was designed to provide comparable information on public and private sectors, as well as trend data over time. The data available through the Annual School Survey is a useful resource to determine overage, enrolment, repetition and dropout rates by gender and province. With the availability and the quality of the data from the Annual Schools Survey key questions can be answered such as: "Where in the system is the highest dropout and repetition?"

Enrolment-driven data management is a central focus of South African government's redress efforts, mainly because of state based accountability policies, such as the South African Schools Act and broader education policy. The aggregated data was always enough to address these policy requirements. The enrolment-driven nature of the education system is reflected in:

- the allocation of school funds based on the National Norms and Standards for School Funding
- the post-provisioning of teachers,
- the grading of primary and secondary schools
- The allocation of management teaching posts, and many more.

Individual-level Data

Unit record (individual) data refers to the data collected for each learner through a school administration and management system (van Wyk & Crouch, 2020). In creating a longitudinal data system, it is necessary to link the different datasets that have been collected for individual learners for each year by using a common field across these datasets.

We need more and better data, because of the focus on equity and equality. Once equity and equality become a concern you need data not only on input, but on outcomes. Data has been used adequately for policy and planning purposes, but not always for management. That requires data disaggregation at a school by school, learner by learner level and more so linking of data sets. The SBA's in SA-SAMS, DDD and LURITS become more and more important and provide such detail. Individual-level data makes it possible to create a longitudinal dataset and determine:

Analytical Power of longitudinal data

With longitudinal data one can determine:

- exactly how many learners of a specific cohort dropped out of the system,
- how many progressed through the system without any repetition and
- how many are still in the system with one or more repetitions.

With the availability of unit-level learner records key questions can be answered such as:

- "What is the profile of the learners who dropped out of the system?" in terms of age, gender, grade, quintile, etc.
- What is the profile of the learners who progressed without any repetition?" in terms of age, gender, quintile, etc.

Progression Power of longitudinal data

Analysing the growth of learners' academic proficiency over time reveals the successes and areas for improvement of subgroups of learners as well as individual learners. With access to longitudinal data, teachers and principals can follow the academic progress of individual learners across grades and even school systems. Learner-level longitudinal assessments are the only means by which academic growth of individual learners can be calculated (The Data Quality Campaign (DQC),2008).

With longitudinal data we can for example, compare the performance of learners from one term to the next. It is expected that learners that perform well in one term's examination usually do well in the next term's examination. This kind of comparison can help to determine if there was a decrease in the performance of learners per subject. In this way we can identify those learners that need support and help.

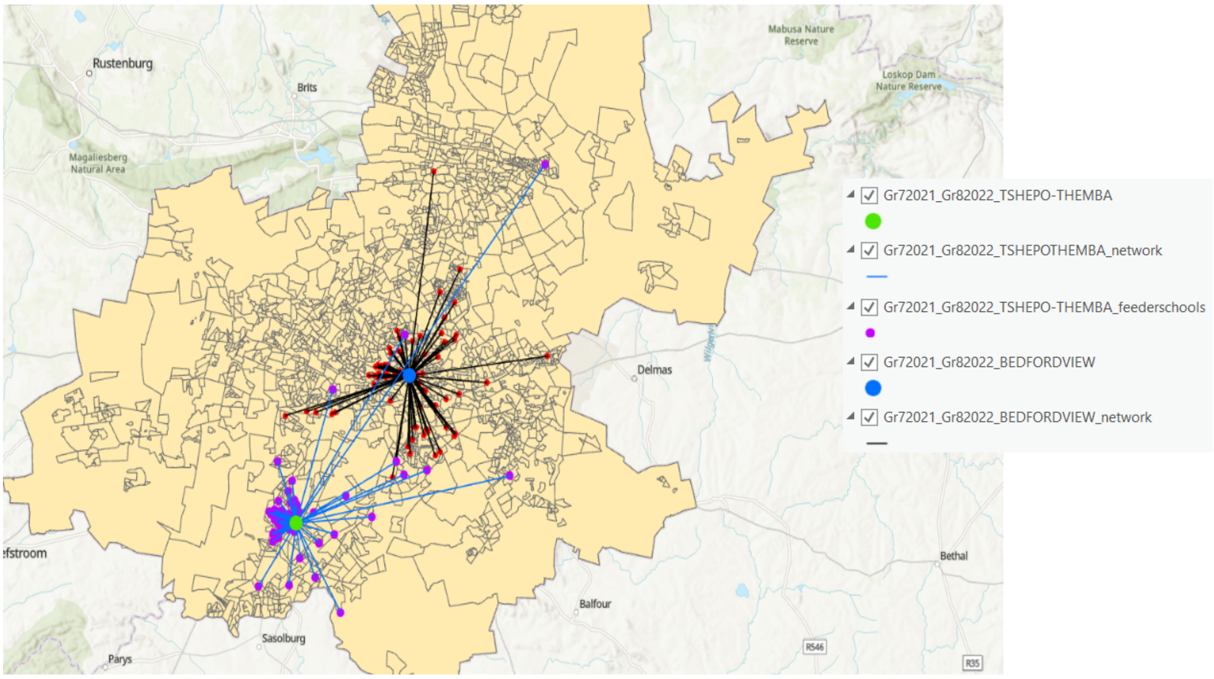
Predictive Power of longitudinal data

With longitudinal data one can for example, determine if the matric 3rd term examination is a predictor for the final examination. Is it that the learners that perform well in the matric 3rd term will also perform well in the NSC final examination? This comparison can be done per subject. This kind of verification also determines if the SBA is at the right level (standard) at a school level.

Mobility of learners

The movement of learners between schools can be determined using longitudinal data. The movement of learners can be result of the transition between primary and secondary schools in terms of school choice in secondary school relative to primary school – feeder schools. With longitudinal data it is possible to determine the primary school of origin of all those learners in the first year of secondary school (Grade 8) when they were in primary school the previous year. For example, **Figure 8** shows the geographical location of learners from the primary schools for a particular secondary school (A secondary school where the feeder schools are 70 plus)

Figure 8: Feeder schools of particular secondary Schools


Rustenburg
Brits
Magaliesberg Natural Area
Mabusa Nature Reserve
Loskop Dam Nature Reserve
Delmas
Wilge
:fstroom
Bethal
Balfour
Sasolburg
Parys
R546
R35
Gr72021_Gr82022_TSHEPO-THEMBA
Gr72021_Gr82022_TSHEPOTHEMBA_network
Gr72021_Gr82022_TSHEPO-THEMBA_feederschools
Gr72021_Gr82022_BEDFORDVIEW
Gr72021_Gr82022_BEDFORDVIEW_network

Source: GDE SA-SAMS database

Cohort Analysis

The availability of individual learner-unit records allows one to track learners as a group or cohort over a specified period. A cohort is a group of learners that share some common characteristic over a period of time. The learner's unique identifier makes it possible to follow learners' progress in the system through the identifier in longitudinal data (data gathered on the same learner from year to year).

In analysing the progress of learners, grade 8 learners of 2017 were considered as a cohort and tracked through the school system by using the SA-SAMS data in Gauteng. Through this cohort analysis we could observe:

- The changing of schools by learners over the period of 5 years
- The progression of learners through the system, for example how many learners progressed without repetition, how many learners progressed with one or more repetitions and how many dropped out of the Gauteng Education System during these 5 years.

Progression of learners through the system

Table 6 gives a summary of how the grade 8 learners of 2017 progressed through the system using longitudinal data.

Table 6: Grade 8 cohort by Grade and Year

Grade Id	2017	2018	2019	2020	2021	2022	Total
Gr8	87 053	12 835	1 163	372	98	34	101 555
Gr9	-	66 342	17 155	1 883	401	107	85 888
Gr10	-	**133**	50 901	27 283	7 533	2 059	87 909
Gr11	-	**67**	**108**	30 325	19 340	6 607	56 447
Gr12	-	**5**	**22**	**360**	27 407	14 891	42 685
Total in school	87 053	79 382	69 349	60 223	54 779	23 698	374 484
Repeaters	**0**	**12835**	**18318**	**29538**	**27372**		
Total Dropout/Unaccounted	**0**	**7671**	**10033**	**9126**	**5444**		
Cumulative Dropout/Unaccounted	**0**	**7671**	**17704**	**26830**	**32274**		

Source: DDD data (Own Calculations)

Table 6 shows the learners that progressed through the system without any repetition as indicated with yellow, learners who progressed through the system with repetition indicated with green and the learners who dropped out of the Gauteng Education system indicated with light green.

Figure 9: Grade progression and enrolment among the 2017 GDE Grade 8 cohort

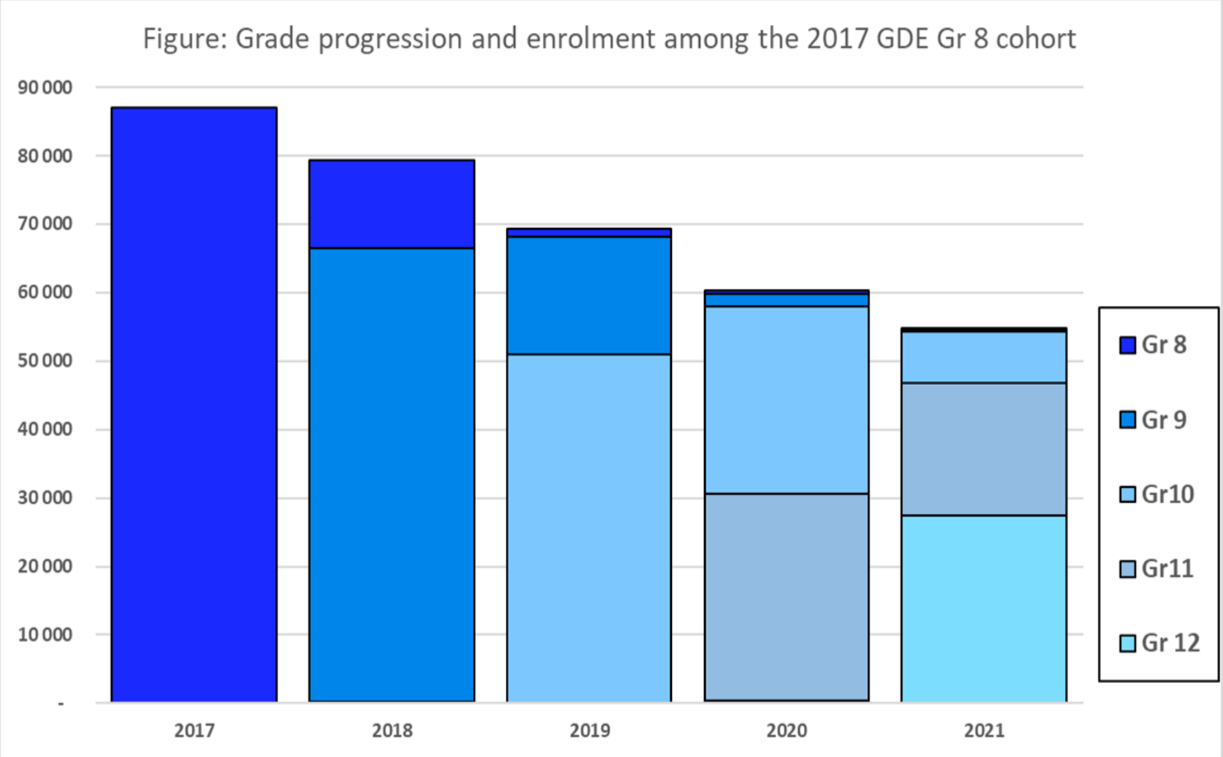
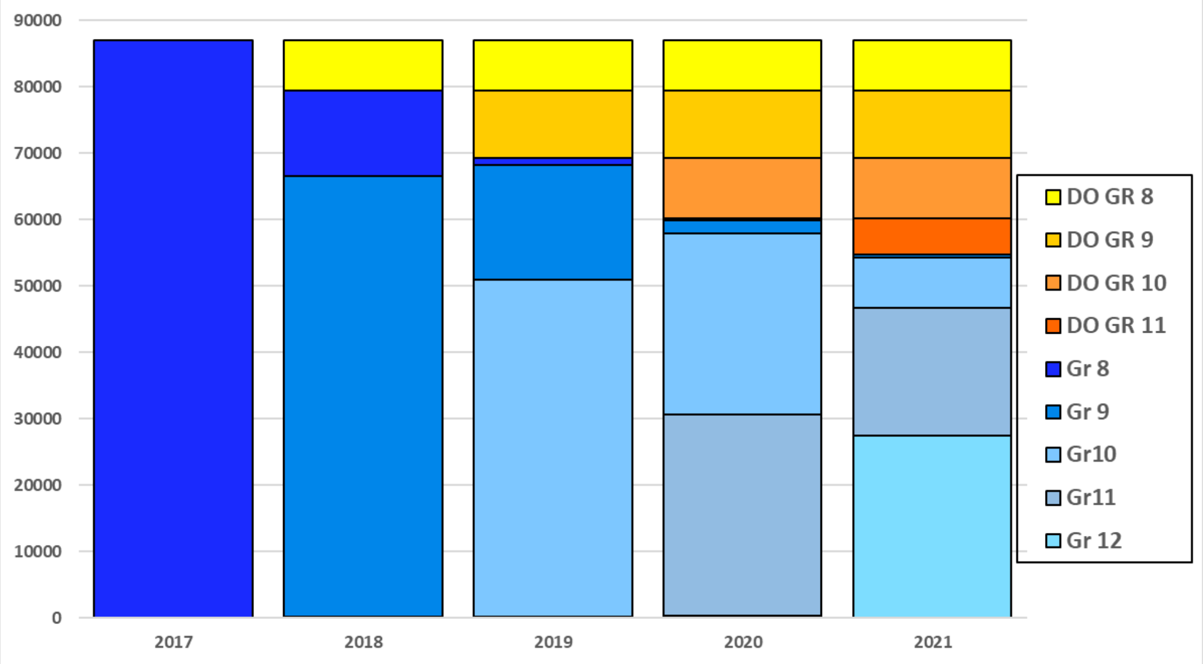


Figure 9 shows how many learners progressed “on track” from Grade 8 to Grade 12 between 2017 and 2021 in GDE schools using DDD data.

Figure 10: Grade progression, enrolment, and drop-out among the 2017 GDE Gr 8 cohort



In addition to the previous information, Figure 10 shows what percentage of learners from the cohort dropped out of the GDE system and when did they drop out.

5. REFERENCES:

CPSI Ltd. 2010. *The Role of the Unique Student Identifier in Longitudinal Data Systems*, CPSI xDUID Unique Identifier System White Paper, Available from: https://www.cpsiltd.com/wp-content/uploads/2015/04/CPSI-LDS-and-the-xDUID.pdf

DBE. 2022. *SA School Administration and Management System*. Retrieved November 2022

The Data Quality Campaign (DQC). 2008. *Tapping into the Power of Longitudinal Data: A Guide for School Leaders*. In partnership with Association of Secondary School Principals, January 2008

Van Wyk C. & Crouch L. 2020. *Efficiency and Effectiveness in Choosing and Using an EMIS: Guidelines for Data Management and Functionality in Education Management Information Systems (EMIS)*. UNESCO, UIS, GPE. Available from: https://tcg.uis.unesco.org/wp-content/uploads/sites/5/2020/09/EMIS-Buyers-Guide-EN-fin-WEB.pdf