Fieldwork Report

Data collection for the Early Grade Reading Study II in Mpumalanga (EGRS II) (Wave 4)

13 December 2019

Submitted by:

Mr Wade Harker
Khulisa Management Services
26 7th Avenue, Parktown North
Johannesburg, 2193, South Africa
Tel: +27 11 447 6464
Fax: +27 11 447 6468
Email: wharker@khulisa.com
www.khulisa.com

Submitted to:

Dr Janeli Kotzé
Department of Basic Education (DBE)
222 Struben Street
Pretoria, South Africa
Email: kotze.j@dbe.gov.za
Cc: Tshegofatso Thulare
Email: Thulare.T@dbe.gov.za

Table of Contents

1 Introduction 5
2 Understanding of the Assignment 5
3 Implementation 6
3.1 Inception 6
3.2 Pilot Testing 7
3.3 Preparation for Fieldwork 7
3.3.1 Fieldworker Screening and recruitment 7
3.3.2 School Mapping 8
3.3.3 Fieldwork Schedule 8
3.3.4 Contacting Schools 9
3.3.5 Pilot Debriefing 9
3.3.6 Fieldworker Manual 9
3.3.7 Tools and Training Materials 9
3.3.8 Fieldworker and Supervisor Training 9
3.3.9 Fieldworker List 11
3.4 Fieldwork 12
3.4.1 Daily fieldwork routine 12
3.4.2 Weekly fieldwork routine 13
3.5 Data Quality Checks 14
3.5.1 Daily data quality checks 14
3.5.1.1 Statistical checks 14
3.5.1.2 Manual checks 14
3.5.2 Weekly data quality checks 15
3.6 Data Preparation and Reporting 15
3.6.1 Learner Assessment Data 16
3.6.2 Contextual questionnaires 16
3.6.3 Data Arrangement 16
3.6.4 Response rates 17
3.6.5 Reasons for incomplete questionnaires 17
4 Challenges, Victories and Lessons 17
4.1.1 Challenges 17
4.1.1.1 Autocomplete and prepopulated fields in data collection instruments 17
4.1.1.2 Daily entry of linking form data 18
4.1.1.3 Availability of the statistician feedback 18
4.1.1.4 Timely daily submissions 18
4.1.1.5 Retrieving school packs from the field 18

4.1.1.6 Tangerine technical glitches 18
4.1.1.7 Translation inconsistencies – learner written assessment 18
4.1.1.8 Updating tangerine during data collection 18
4.1.1.9 Extreme hot weather – fieldworker fatigue 19
4.1.1.10 High number of learner transfers 19
4.1.1.11 Fieldworker illness 19
4.1.2 Victories 19
4.1.2.1 Completing data collection timeously 19
4.1.2.2 Fieldworker and supervisor recruitment 19
4.1.2.3 Administration of exam-style fieldworker competency assessment 19
4.1.2.4 Working relationship with the DBE 19
4.1.3 Lessons and Recommendations 19
5 Project Team 20
Annex 1. Agreed Data Cleaning Processes 23
Annex 2. Data Quality Checks Report 24
Annex 3. Data Quality Assessment and Surveillance Plan (QASP) 34

List of Tables

Table 1: Fieldworker and Supervisor Selection Criteria 7
Table 2: Data Cleaning Interpretation Legend 16

List of Figures

Figure 1 Four-phase approach 6
Figure 2 Spatial distribution of EGRS II data collection sites 8
Figure 3: Core Project Team Organogram 20

1 Introduction

In 2019, Khulisa Management Services (Pty) Ltd (hereinafter referred to as "Khulisa") was reappointed to provide data collection services to the South African Department of Basic Education's (DBE) Early Grade Reading Study II (EGRS II) in Mpumalanga – the fourth wave of data collection for the study (Wave 4). The services rendered comprise the collection, quality assurance, and submission of learner assessment data and contextual data in 180 schools in the province.

The EGRS II is a United States Agency for International Development (USAID)-funded project run by the DBE in collaboration with the University of the Witwatersrand (WITS) School of Education. The project focuses on providing support to teachers in teaching English First Additional Language (EFAL) in the Foundation Phase, leading to improved teacher practices in the classroom and, ultimately, to improved learner performance.

Khulisa is pleased to present this **Fieldwork Report** for the Wave 4 data collection for the DBE's evaluation of the EGRS II. This report describes Khulisa's understanding of the assignment and the approach implemented. The report detail's the data collection protocol followed, the response rates, the challenges experienced, as well as lessons learned.

2 Understanding of the Assignment

It is well documented that the South African education system is facing numerous systemic or "binding constraints"[1]. This is particularly evident at the Foundation Phase, where learners struggle to read in their home language, which has a knock on effect throughout their future learning and development. The DBE, through EGRS II, seeks to expand its research into alternative approaches to strengthening teaching and learning in the Foundation Phase.

The purpose of this assignment was to support the DBE in evaluating the EGRS II by collecting data from 180 schools across two Mpumalanga districts: Gert Sibande and Ehlanzeni, as part of the EGRS II Wave 4 assessment. Thirty fieldworkers were paired-up (i.e. 15 groups of two fieldworkers each), and each pair spent a full day in each school to administer the following instruments:

- Learner Oral Assessment (LA)
- Learner Written Assessment (LWA)
- Principal Questionnaire (PQ)
- School Observation (SO)
- Teacher Questionnaire (TQ)
- Linking Form (LF)

In addition, Khulisa was responsible for training, fielding, and managing the team of 30 fieldworkers as well as training, fielding, and managing five supervisors. Supervisors were the first point in a 3-tier quality assurance strategy.

The DBE Project Management Team (PMT) was responsible for developing and pilot testing all research tools. Khulisa made provision for the DBE to develop all software-based tools on the Khulisa cloud to aid the turnaround of tool revisions.

Upon completion of fieldwork, Khulisa engaged in data preparation and reporting. Data preparation

[1] Van der Berg, S. et al. (2016). Identifying Binding Constraints in Education. RESEP.

and reporting included a number of activities, listed below:

- A post-fieldwork data preparation meeting;
- Basic cleaning of data collected during the EGRS Wave 4 data collection assignment[2]; and
- Submission of datasets and reports to the DBE PMT by 10 December 2019.

3 Implementation

It is Khulisa's standard practice to establish a participative and consultative working relationship with the client around the task to be completed. Khulisa has worked collaboratively with the DBE since 2017, when Khulisa was initially appointed as a service provider to administer the EGRS II Wave 2 data collection. During this assignment, Khulisa communicated on a regular basis with the DBE PMT, particularly in cases where challenges observed could substantially affect the results of the study. This hands-on approach between the DBE and Khulisa supported the successful administration of the EGRS II Wave 4 data collection.

Khulisa employed a 4-phase approach to this assignment (see Figure 1). Although structured chronologically, some of the activities overlapped between the phases.

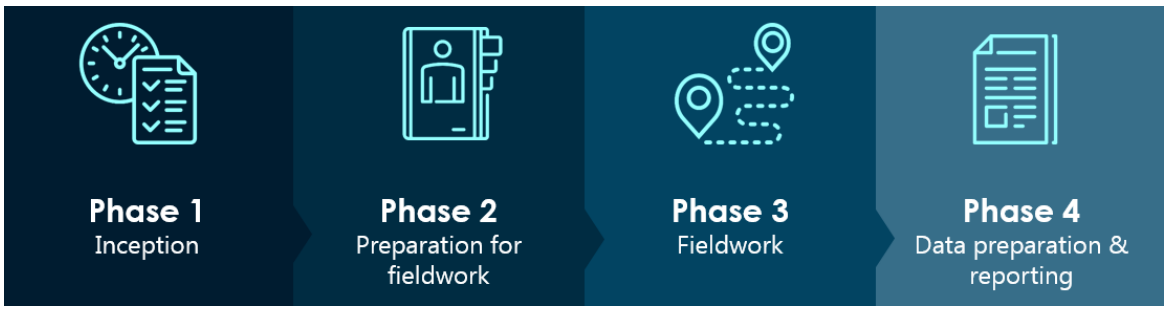


Figure 1 Four-phase approach

3.1 Inception

The inception meeting for this assignment was held at the DBE offices in Pretoria on Monday, 16 September 2019. USAID, DBE, and Khulisa representatives attended the inception meeting. Dr Janeli Kotze led the meeting and was supported by Ms Tshegofatso Thulare (DBE). Although not budgeted, Khulisa brought on board a statistician to support the data quality objectives of the assignment. The statistician, Dr Petra Gaylard, also attended the inception meeting.

Mr Wade Harker (Project Manager) and Ms Katharine Tjasink (Senior Technical Oversight) provided a summary of Khulisa's approach to the 2019 EGRS II Wave 4 data collection in Mpumalanga. Mr Harker shared lessons learned from previous waves of the EGRS, as well as insight from the implementation of other relevant data collection assignments, and proposed new strategies to improve data quality and monitoring.

The stakeholders welcomed the suggestions and agreed that adequate reinforcements to data quality will be implemented. The participants discussed next steps and agreed on submission dates for key deliverables.

[2] A list of agreed data cleaning processes is provided in Annex 1

3.2 Pilot Testing

During the reappointment negotiations, the DBE, WHC, and Khulisa explored various implementation options. The DBE PMT decided to run the pilot test independently, with Khulisa providing support to the DBE where needed. As such, Khulisa was not part of the pilot testing process. However, the Khulisa team engaged in multiple post-pilot debriefing meetings. Key points discussed in these meetings include the following:

- Lessons learned from the pilot testing process and the implications for learner assessment tool review;
- Changes required for each of the research tools;
- Fieldworker requirements and recruitment;
- Review of quality assurance procedures and tools;
- Fieldwork logistics planning;
- Tablet preparation and cloud backend reviews; and
- Contacting schools and sharing information regarding the anticipated data collection.

Khulisa provided ad hoc support to the DBE PMT during the pilot testing process. In addition, Khulisa and the DBE PMT discussed options as to how the fieldworker competency assessment should be administered. The two parties agreed that the fieldworker trainees would be assessed in an examination style, and that video recordings of learner assessments would be used as part of the fieldworker competency assessment material.

3.3 Preparation for Fieldwork

3.3.1 Fieldworker Screening and Recruitment

Khulisa started early preparations for fieldwork during the inception phase. This included screening and recruitment of fieldworkers. The screening process comprised background checks, verification of qualifications, as well as a telephonic interview. Based on the lessons learned from previous years (2017 and 2018), the average fieldworker age was decreased. This is due to the physical requirements and productivity expected of the fieldworker cohort.

Towards the end of the fieldworker screening process, Khulisa shortlisted a group of 36 fieldworkers and 5 fieldwork supervisors to attend the EGRS II Wave 4 Fieldworker Training in Mpumalanga. The fieldworker group comprised an even split of fieldworkers who had been previously involved in EGRS and fieldworkers who had not been involved in any EGRS work. The group of fieldwork supervisors were all involved in previous EGRS data collection assignments.

Table 1: Fieldworker and Supervisor Selection Criteria

Fieldworker Selection Criteria	Fieldwork Supervisor Selection Criteria
1. Degree/teaching diploma OR Matric with teaching experience	1. Degree and/or teaching experience
2. Researcher work experience – previous learner assessment experience advantageous.	2. At Least 3 years work experience in the education sector – previous EGRS experience (Mandatory)
3. Fluency in English, SiSwati and/or isiZulu languages	3. Fluency in English, SiSwati and/or isiZulu languages
4. Previous experience working with foundation phase learners advantageous	4. Previous experience working with foundation phase learners (mandatory)

5. Driver's license (advantageous)	5. Driver's license (Mandatory)
6. Good interpersonal skills	6. Good interpersonal skills
	7. Experience managing fieldworkers

Overall, the strategic decision to recruit a younger group of fieldworkers has paid-off as a marked improvement in fieldworker productivity was observed throughout the EGRS II Wave 4 data collection process.

3.3.2 School Mapping

To improve fieldwork efficiency, Khulisa plotted GPS coordinates provided by the DBE PMT to form fieldwork clusters in each of the two districts. Minor changes were made to the fieldwork clusters used in 2017 and 2018. In a handful of schools, the GPS locations were inaccurate. However, this did not affect fieldwork teams in reaching their destinations. Overall, using mapping and school clustering improved fieldwork planning and logistics.

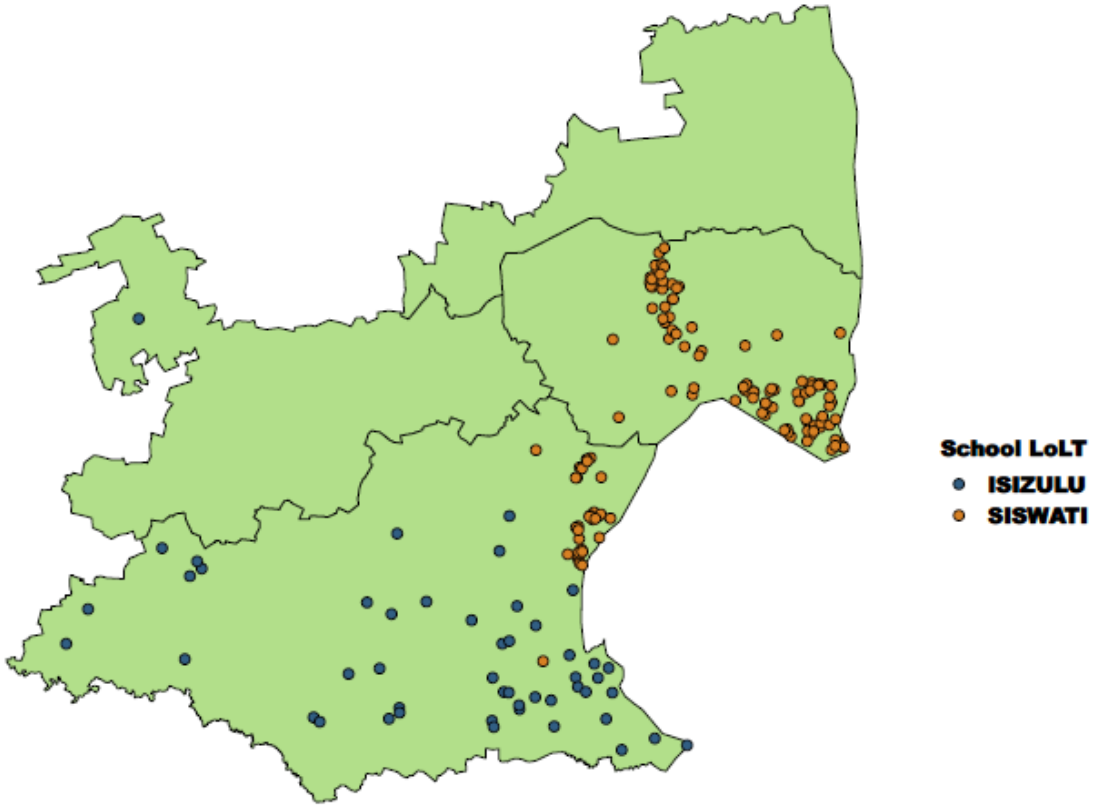


Figure 2 Spatial distribution of EGRS II data collection sites

3.3.3 Fieldwork Schedule

Using the school clusters mapped in Google maps, Khulisa developed a draft fieldwork schedule as an initial step. Thereafter, Khulisa secured visitations with each school as per the draft fieldwork schedule. The fieldwork schedule indicated which schools were Khulisa sample schools and which schools were vocabulary-testing schools[3]. Additionally, Khulisa reviewed the fieldwork schedule to ensure that each fieldworker team had a relatively equal split of treatment and control schools. Fifteen fieldwork teams

[3] The DBE selected 60 schools in the sample to participate in a retest and a Productive Vocabulary Test (PVT). The DBE carried this out independently.

were assigned to assess 12 of the 180 sample schools. Although it was not a stated aim of the project to randomise fieldworker teams to school groups, it was nonetheless preferable that there should not be a gross imbalance between fieldworker team and group, which could lead to bias.

Khulisa cross-tabulated fieldworker teams and school groups. A X^2 test was used to assess the relationship between these two variables. The association between team and school group was not significant (p=0.48). The results showed no significant bias between fieldworker team allocation and school group.

Khulisa worked closely with the DBE to ensure that the fieldwork schedule accommodated both the DBE and Khulisa's assumed logistics and responsibilities.

3.3.4 Contacting Schools

Khulisa contacted all 180 sample schools prior to fieldwork to provide detailed information about when schools would be visited. Schools were first contacted via telephone, followed by an email detailing the anticipated series of events at the schools. This approach formed part of a multi-stage learner tracking strategy. Figure 2 depicts the spatial distribution of the approximate data collection sites to be visited during the EGRS II Wave 4 data collection.

3.3.5 Pilot Debriefing

On 08 October 2019, Khulisa met with the DBE to discuss the DBE's pilot observations and lessons learned. The DBE PMT explained the protocol followed during the pilot testing process as well as the tools tested. The DBE PMT advised that only the learner assessment tools were pilot tested, because the contextual tools (principal and teacher questionnaire) were used during three previous waves of data collection and therefore required minimal modification. Khulisa discussed and agreed the planned logistics as well as the training approach with the DBE PMT.

3.3.6 Fieldworker Manual

The DBE PMT lead trainer developed a training manual. Khulisa did not develop an additional manual, as the DBE training manual was sufficiently comprehensive.

3.3.7 Tools and Training Materials

Following the pilot debriefing, and after the DBE PMT finalised the research tools, Khulisa printed and packaged all paper-based training tools. The paper-based tools as well as the data collection tablets were delivered to the training venue in Mpumalanga. The data collection tablets were programmed for use the night before the fieldworker training commenced.

3.3.8 Fieldworker and Supervisor Training

Khulisa hosted a five-day training workshop (14 to 18 October 2019) at a training venue in Mpumalanga, led by the DBE lead trainer Mrs Maxine Schaefer and co-facilitated by Khulisa Project Manager Mr Wade Harker. The Khulisa team, DBE PMT, EGRS tool developer as well as the shortlisted supervisors and fieldworkers attended the training workshop.

Per standard procedure, Khulisa trained a larger group of candidate fieldworkers (36), from which the final group of fieldworkers (30) were selected based on their in-training performance and the results of a fieldworker competency assessment administered by the DBE PMT. Khulisa preselected five fieldwork supervisors who had been part of the previous EGRS waves of data collection.

For the first two days of training, the lead trainer focussed on the learner assessment tools. This process involved a task-by-task training approach, which ensured that the fieldworkers gained a solid understanding. Additionally, the lead trainer emphasised speed and accuracy of delivery of the learner assessment tools.

Fieldwork supervisors were utilized to oversee breakaway sessions, observe fieldworker performance, and identify fieldworkers who required intensive support. The lead trainer trained fieldworkers identified as "require intensive support" separately, while the designated fieldwork supervisors observed the rest of the fieldworkers. Thus, intense training and role-play methods were used during the first two days of fieldworker training.

During the last session of day two, fieldworkers were allocated to school simulation groups. Each fieldwork supervisor was assigned to a cohort of fieldworkers. The five designated supervisors were responsible for the supervision and recording of fieldworker performance during the in-school simulation experiential learning activity.

On day three (in-school simulation day), five teams of fieldworkers were sent to five schools - one school per team. Two of the schools were located in Gert Sibande district (isiZulu schools) and three schools were located in Ehlanzeni district (SiSwati schools). The teams were divided according to language proficiency. The two fieldworker teams allocated to Gert Sibande left at 06:30am, as they had to travel more than two hours to reach their simulation schools. The remaining three fieldworker teams allocated to Ehlanzeni departed from the training venue at 07:30am, as the journey to their respective simulation schools was relative short in comparison to the Gert Sibande teams. In-school simulation was expected to commence by 09:00am.

In-school simulation entailed the following:

- Fieldworkers practiced the expected daily routine;
- Fieldworkers practiced in-school introduction protocols;
- Fieldworkers practiced setting up assessment locations;
- Fieldworkers obtained relevant documentation such as class lists;
- Fieldworkers practiced the administration of learner assessments in a real-life situation;
- Fieldworkers improved their proficiency and accuracy using tablets;
- Supervisors practised supervisory protocols and identified mistakes; and
- Supervisors familiarised themselves with the relevant quality protocols as well as their roles and responsibilities.

The simulation day was scheduled to end at 13:00pm on day 3. Thereafter, all teams were to return to the training venue in Nelspruit. Due to the anticipated late return of the two Gert Sibande teams, the lead trainer administered a post-simulation debrief session with the three Ehlanzeni groups at 15:00pm. The debrief addressed the challenges faced, as well as the feedback provided by the fieldwork supervisors. The lead researcher administered this debriefing session within an hour and made herself available a one-on-one question and answer session afterwards. The two Gert Sibande teams arrived at Nelspruit at approximately 15:30pm. Visibly exhausted, the Khulisa Project Manager advised them to have lunch and then rest, with the understanding that a simulation debriefing session would be administered later that evening. The two Gert Sibande teams were debriefed at approximately 20:00pm. The relevant tool-specific feedback retrieved from all groups was incorporated accordingly by the DBE on the same day.

On the fourth morning of fieldworker training, the team discussed the lessons learned before undergoing a fieldworker competency test. The DBE PMT led this process. The conditions for this competency assessment were as follows:

- Exam style conditions'

- Room arranged to suite the activity at hand;
- Audio and visual quality confirmed with trainees;
- No clarification of questions allowed during the process;
- Two rounds of assessments;
- For each round, video clips of each learner assessment task were played and projected on the large boardroom screen;
- Fieldworkers selected responses as per the video clips; and
- In between each task, a window of up to two minutes was granted to the fieldworkers to make any notes.

Upon completion of the fieldworker competency assessments, the Khulisa Project Manager submitted the data in conjunction with the fieldworkers. All assessment data was submitted to the Khulisa cloud. Thereafter, Khulisa and DBE PMT trained the fieldworkers on the contextual research tools (principal questionnaire, teacher questionnaire, and linking forms). The facilitators trained the fieldworkers in a question-by-question manner. Further breakaway and roleplay sessions continued until the end of day four.

While the breakaway sessions were occurring, a DBE representative analysed the fieldworker competency assessment data. The analysis was shared with the Khulisa Project Manager. The results showed that, out of the 36 fieldworkers assessed, six were in need of intensive support. These six fieldworkers were supported by the supervisors who arranged after hours support sessions.

On the fifth day, the team held a final technical debriefing session to address all inconsistencies. For practical reasons, it was suggested that the principal survey be split into two separate tools because the survey featured interview questions as well as school terrain observations. All other relevant suggestions were considered during the DBE PMT's final review of the data collection tools.

Thereafter, the Khulisa PMT focussed on the fieldwork administration and logistical arrangements. The key activities of the logistics session are summarised as follows:

- The final fieldworkers were selected;
- Fieldworkers were assigned to teams;
- Fieldworkers were assigned to supervisors;
- Fieldworkers were shown their enumerator areas using Khulisa Geographical Information System (GIS) data;
- Fieldwork accommodation, transport and associated logistics were discussed; and
- An overview of the Standard Operating Procedures as well as roles and responsibilities were discussed.

After all logistics were covered, the fieldworker training was adjourned at 16:00pm on Friday 18 October 2019.

3.3.9 Fieldworker List

A final fieldworker list was sent to the DBE on 25 October 2019. As explained above, a series of protocols were considered during the selection of the final group of fieldworkers.

3.4 Fieldwork

Fieldwork took place from 28 October to 15 November 2019 in 180 schools. Fifteen fieldwork teams (two fieldworkers per team) were deployed to collect the necessary data. At each school, Khulisa collected data from a maximum of 20 Grade 3 learners (some in other grades), Grade 3 teachers, and school principals. Fieldworkers were also tasked with assessing the sample of learners who had repeated Grade 1 or 2, in 2017 and/or 2018.

3.4.1 Daily fieldwork routine

Fieldworkers were responsible for the administration of various activities at each school on a daily basis. These included:

- Preparing all data collection materials and instruments ahead of the school visit.
- Arriving at the school by 07:30am.
- Taking a photo of the school name board and sending it via their respective supervisory WhatsApp groups (Khulisa check-in mechanism)
- Introducing the fieldworker team to the school management team.
- Setting up the assessment room for learner assessments.
- Retrieving all learner and teacher identifying data (class lists) to assist the team with learner and teacher identification.
- Assessing up to 20 sample learners (mostly in Grade 3).
- Interviewing all Grade 3 teachers responsible for teaching the 20 sample learners.
- Observing the classroom of all interviewed Grade 3 teachers.
- Administering a short teacher exercise with all interviewed Grade 3 teachers.
- Observing the DBE workbook of one proficient learner for each of the interviewed Grade 3 teachers.
- Interviewing the school principal.
- Accurately filling in the learner linking form.
- Capturing up to 20 learner written assessments on Tangerine after school.
- Sending school check-in photo via WhatsApp by 14:30pm.
- Sending photos of the learner linking form via WhatsApp by 14:30pm.
- Sending photos of all relevant class lists via WhatsApp by 14:30pm.
- Sending a photo of the school checklist via WhatsApp by 14:30pm.
- Synchronizing all data collected to the Khulisa cloud via Tangerine by 17:00pm.

Apart from the team in the field, Khulisa had three other technical project resources dedicated to data quality checks. Their responsibilities were:

1. **Data Capturer** – capturing of all information shared via WhatsApp, including the learner linking forms, class lists, and checklists. Any discrepancies observed were communicated directly with the relevant fieldwork supervisor for correction.
2. **Data Quality Monitor** – administering various quality protocols on the actual physical data packs returning from the field. This resource closely monitored completeness of the linking

forms vs the class lists collected; the number of written assessments vs the oral assessments; the number of teacher questionnaires vs the linking forms and class lists as well as the workbook data for consistency. As part of this process, the quality assessor could identify all make-up requirements on a weekly basis, barring the first week of data collection.

3. **Backend Data Quality Assessor** – conducting daily high frequency checks. A set of criteria were decided and agreed upon between Khulisa and the DBE PMT prior to data collection. This process involved extracting data from the Tangerine backend, and performing various statistical checks on the raw data. The results were shared with the DBE.

Overall, the daily routines were adhered to fairly well apart from a couple of cases where submissions were delayed due to unforeseen circumstances.

3.4.2 Weekly fieldwork routine

Primary data collection took place from Monday to Thursday each week. Fridays were dedicated to fieldwork make-up, which entailed collecting data that could not be collected during the Monday to Thursday data collection window.

As noted earlier in this report, fifteen pairs of fieldworkers carried out primary data collection over a 3-week period. Each fieldworker in the pair was equally responsible for executing their allotted work.

Prior to fieldwork, Khulisa developed daily fieldwork schedules and learner linking forms to identify the learners to be assessed. Fieldworkers used the learner linking information provided to select the correct learners. Based on previous experience with school and learner assessments, we understood that the 'typical school day' was highly susceptible to change and disruptions. As such, fieldworkers used the daily fieldwork schedule as an illustrative guide for completing data collection. Upon completion of fieldwork activities at each school, fieldworkers obtained an official school-stamped document as evidence of the school visit.

Each fieldwork team was assigned 12 schools for the duration of the data collection. Thus, the weekly data collection arrangement included four days data collection and one day of make-up data collection each week. This approach worked well as it allowed teams who could not complete their allotted work during the week to complete it on Fridays.

Fieldwork teams were assigned to fieldwork clusters based on their language of proficiency to ensure that all research tools were administered accurately. Both fieldworkers in each team were fluent in the school's language of learning and teaching (LoLT). Using their fieldwork schedules, fieldwork teams contacted their allocated schools a day prior to the scheduled visit to ensure that the relevant schools expected the visit. In some cases (five), principals indicated that the school management team had not been made aware of the visitation. However, Khulisa had multiple engagements with each school notwithstanding the fact that many schools were very hard to reach.

As noted earlier in this report, during the EGRS II Wave 4 data collection the DBE PMT deployed a smaller research team who were mandated to conduct vocabulary assessments (hereafter referred to as DBE vocab team) in 60 of the 180 sample schools. This arrangement required the Khulisa team to coordinate and work with the DBE vocab team to ensure that the learners identified as vocab learners be prioritised for assessments early in the day to allow the vocab team to complete their allotted work timeously. Although some coordination challenges were observed, both groups of researchers could generally execute their work in the given timeframes.

As described earlier, Khulisa deployed five fieldwork supervisors strategically for the duration of the data collection. One fieldwork supervisor was responsible for three fieldwork teams (two researchers per team). Supervisors provided ongoing support to fieldworkers, and served as an initial point of

contact for any challenges encountered in the field. The supervisors were trained with the fieldworkers to ensure that they were able to administer all research instruments should the need arise.

During the first three days of data collection, supervisors monitored their teams intensely. This allowed supervisors to quickly identify and address data quality concerns. Supervisors had check-in meetings with the Khulisa Project Coordinator and Project Manager in the evening of each data collection day to discuss issues, victories, and actions.

3.5 Data Quality Checks

3.5.1 DAILY DATA QUALITY CHECKS

3.5.1.1 Statistical checks

On each day following a day of main fieldwork (excluding Friday catch-up days), all the learner oral assessment, learner written assessment, teacher interview, principal interview, school observation and linking files were submitted for evaluation, together with the fieldwork schedule and master file of learners to be evaluated received at the start of fieldwork.

Data quality checks were carried out according to a list of checks agreed with the client prior to the start of fieldwork. The checks were programmed in Stata. Results were output to an HTML file and the salient findings were also summarised in an Excel file ("EGRS data checks tracking Day XX"), both of which were sent by the statistician to the Khulisa Technical Representative and the Fieldwork Provider, on the same day.

3.5.1.2 Manual checks

In addition to the statistical checks, the following manual capture and checks were done on a daily basis:

- Capture the linking forms and other accompanying documentation for each team on a daily basis.
- Check that all schools assessed were on the school list for the day
- Check the number of principal consent forms and Grade 3 teacher consent forms captured
- Check the number of class lists submitted versus required
- Check the number of Grade 3 teacher exercises and observations submitted
- Check the number of Grade 3 learner workbooks assessed
- Check that the Grade 3 learner names in the workbooks and the ID captured on the linking form match
- Check the number of learners assessed versus the number of learners on the linking forms
- Check and record the number of learners that transferred
- Check and record the number of learners that transferred
- Check and record the number of learners assessed
- Capture/Mark the learner written assessment booklets before 17:00

The manual checks were a crucial part of the day-to-day fieldwork activity. As described above, fieldworkers were required to perform various checks. Before leaving the school visited for the day, fieldworkers were to complete all checks baring the capturing of the learner assessment which had to be completed by 17:00. Supervisors were responsible for checking in with each of their allocated teams before submitting their data for the day. All electronic data submissions were due 17:00 of each

day.

3.5.2 WEEKLY DATA QUALITY CHECKS

Each week following a full week of fieldwork, the following investigations were done using all data collected in the linking file, learner oral assessment, and learner written assessment from the Monday to Sunday of the previous week:

- The number of school and number of learners to be assessed was determined from the linking file.
- Using the "Comments" field in the linking file, learners who were marked absent, ill, transferred, or untraceable were removed from consideration.
- The learner oral assessment data file was stripped of records that were empty, started before 07:30, and were duplicate linking IDs with missing data. Remaining duplicates (by linking ID) were documented.
- Next, the learner oral assessment file and linking file were merged by linking ID.
    - Mismatches due to linking ID mistyping were corrected by hand as far as possible.
    - The number of learners in the linking form, but with no oral assessment, was documented.
    - The number of learners with oral assessments who were not in the linking file, was documented, and reasons for the existence of these records were sought.
    - Next, among the records that matched directly on linking ID, the number of true teacher name mismatches (excluding spelling errors, inclusion of second name, etc.) was determined.
- The learner written assessment data file was stripped of records that were empty and were duplicate linking IDs with missing data. Remaining duplicates (by linking ID) were documented.
- Next, the learner oral and written assessment files were merged by linking ID.
    - Mismatches due to linking ID mistyping were corrected by hand as far as possible.
    - The number of learners in the linking form, but with no oral assessment, was documented.
    - The number of learners with written assessments who were not in the oral assessment file, was documented, and reasons for the existence of these records were sought.

The checks were programmed in Stata. The findings were summarised in an Excel file ("EGRS investigations Week X"), which was sent by the statistician to the Khulisa Technical Representative and the Fieldwork Provider.

3.6 Data Preparation and Reporting

The data preparation phase commenced immediately after Khulisa and the DBE held a post fieldwork data cleaning responsibilities meeting. A data cleaning approach was agreed upon by Khulisa and the DBE, **see Annex 1.** Firstly, Khulisa prioritised establishing the return rates for each research tool to ensure that all data has been collected. Once the approximate return rates were established, Khulisa prioritised the capturing and finalisation of the electronic linking database. The linking database was used as the "reference database" to ensure that, school, teacher and learner identifying information is captured consistently across all datasets. Once the linking database were deemed ready to be used for cross checking, the data cleaning team focused on the learner assessment data.

3.6.1 Learner Assessment Data

The learner assessment data were stored in two different datasets per language:

1. Learner Oral Assessment; and
2. Learner Written Assessment.

The team started the learner data cleaning procedure by cleaning the "isiZulu oral" database and thereafter moved over to the "isiZulu written Assessment". A similar process was followed for the SiSwati learner data. The cleaning involved reviewing and rectifying the following information;

- School name
- School EMIS number
- Learner Name and Surname
- Learner Unique ID
- Learner Grade
- Teacher Name and Surname

As per our agreement with the DBE, Khulisa inserted extra columns to showcase the corrected information for each of the abovementioned variables. These extra variables/columns are highlighted in "Green". Instead of adding comments only, Khulisa developed a colour-coded editing criteria which will guide the data interpreter. The criteria is as follows:

Table 2: Data Cleaning Interpretation Legend

	Corrected columns - recommend use
	Duplicate, please investigate
	Blank row, partially blank row, practice data - recommend remove
	Wrongly captured - recommend remove

Khulisa developed this criteria in the hope that it would assist the data interpreter to easily identify corrected information, duplicate information, incomplete information, and data which was wrongly captured. It must be noted, that rows indicated as partially completed are "tangerine technical software related". However, for those "incomplete cases", complete associated data do exist within the datasets.

3.6.2 Contextual questionnaires

Three contextual questionnaires were administered during the data collection period, namely;

- Teacher questionnaire,
- Principal questionnaire
- School observation

3.6.3 Data Arrangement

The same data cleaning criteria was used for all datasets. Each dataset has been arranged similarly. Once each excel workbook has been opened, the user will have access to the following:

1. Sheet 1: Introduction page and the Data Cleaning Interpretation Legend
2. Sheet 2: Cleaned data sorted to the top and duplicates as well as potentially removable data sorted to the bottom of the sheet.

The general thought was to arrange the data sheet in such a way, that the user immediately access the cleaned and corrected data first (on top) and then, if the need arise, investigate duplicate entries as well as potential removable data towards the end (bottom) of each dataset.

3.6.4 Response rates

The response rates were calculated

Research tool	Total Expected	Total Collected	Response Rate
Learner Oral Assessment	3327 (Wave 1 - 2016)	2694 (Wave 4 - 2019)	81%
Learner Written Assessment	2694	2661	99%
Teacher Questionnaire	NA	266	NA
Principal Questionnaire	180	180	100%
School Observation	180	179	99%

3.6.5 Reasons for incomplete questionnaires

Research tool	Reasons
Learner Oral Assessment	High transfer rates coupled with absenteeism
Learner Written Assessment	Learner became unavailable or could not be tracked towards the end of the school day. The learner written assessments were general scheduled to be administered from12:30 each day.
Teacher Questionnaire	We have observed some teacher absenteeism and a small number of cases where teachers were on extended sick leave.
Principal Questionnaire	NA
School Observation	We could not administer 1 school observation due to school logistical constraints.

4 Challenges, Victories and Lessons

4.1.1 Challenges

4.1.1.1 Autocomplete and prepopulated fields in data collection instruments

Prior to fieldwork, the Khulisa team flagged a number of issues with the DBE PMT, which the statistician flagged as issues in the previous EGRS II Wave 3 dataset. These include:

- That the enumerator should be a coded field with a drop-down list. The Wave 3 dataset showed that people spell their names in multiple different ways.
- That the school name should be a coded field (auto-complete) for the same reason listed above.
- That the school EMIS number format should be fixed to force entry of the correct number of digits.
- That the age, number of years, number of learners, number of days etc. should be forced numeric entries ONLY plus known codes for "don't know" (e.g. 98). For example, this would prevent the entry of "2 years" instead of "2", or "O" in place of "0".

The DBE PMT responded that it would not be possible to create autocomplete or prepopulated fields for any of the instruments. The rationale was that since all linking forms for fieldworkers would be

prepopulated with EMIS and school name information clearly identified, this would reduce inconsistencies in capturing the information on assessments and questionnaires.

The daily data quality checks revealed that these flagged issues were not resolved via the use of linking forms. The findings suggest that these issues need to be revisited in future data collection for the EGRS II.

4.1.1.2 Daily entry of linking form data

The daily office-based capturing of linking forms were introduced during the 2019 EGRS data collection process as part of our improvement strategy. Although it added value, and aided the team identifying linking form capturing inconsistencies, we found it challenging to capture linking forms for 15 teams before close of business each day.

4.1.1.3 Availability of the statistician feedback

The availability of the statistician feedback was dependant on the linking forms captured each day. In order to perform the requisite quality checks, the statistician made use of the daily captured linking form. Thus, any delays experienced while capturing linking data would directly affect the availability of our statistician's feedback.

4.1.1.4 Timely daily submissions

During week one, some research teams did not submit their data timeously, which delayed data extraction processes at the end of the day.

4.1.1.5 Retrieving school packs from the field

Retrieving school packs from the field could only be done at the end of each week due to wide geographic spread of each team. As such, it didn't allow the team to attend to discrepancies as timeously as needed.

4.1.1.6 Tangerine technical glitches

The tablets used for data collection had been used various occasions over the past 3-4 years. Based on the field reports, numerous incidences of tablets stalling/freezing during use was reported. The performance of the tablets were also impacted upon by the extreme hot weather experienced during the data collection period.

4.1.1.7 Translation inconsistencies – learner written assessment

During week 1 of data collection, one of the supervisors reported inconsistent translation in one passage in the learner written assessment booklet. This issue was reported to the DBE and a resolution was communicated to the Khulisa Project Manager. Khulisa implemented the recommended solution provided by the DBE, which involved reprinting the affected page and replacing it with the correct version in each SiSwati Learner Written Assessment booklet.

4.1.1.8 Updating tangerine during data collection

During week 1 of data collection, we discovered a software glitch in the teacher questionnaire. The glitch caused the fieldworkers not to progress beyond a certain point in the teacher questionnaire. Khulisa reported this issue to the DBE and the DBE swiftly made the necessary change in the Tangerine backend. The change was a functionality change and not a content change.

As a result, Khulisa had to re-download an updated version of tangerine on all 30 fieldwork tablets in order for the functionality issue to be resolved. The project coordinator and supervisors were prioritised to download the latest version onto all 30 fieldworker tablets. Logistically, it was not feasible to physically go to each fieldworker due to their geographic spread. Thus the instructions for

the updates were done telephonically by the fieldwork supervisors and project coordinator. During this process, two tablets were not updated accordingly. These two affected tablets had the final versions of the research tools, barring the one technical error (as explained above). As such, the data for these tablets were redirected to a separate data collection group. However, because final changes for fieldwork were done on the 19th of October (day after training concluded), the data stored in this separate data collection group was useable. We included the data into the relevant final datasets and attached the associated raw additional datasets as a reference.

4.1.1.9 Extreme hot weather – fieldworker fatigue

The extreme heat experienced during data collection caused fatigue amongst the fieldworkers.

4.1.1.10 High number of learner transfers

Although uncontrollable, the team observed a high number transfers. The unofficial feedback from the field is that, many learner transfers are due to parent migrant labour.

4.1.1.11 Fieldworker illness

One fieldwork fell seriously ill during data collection. His illness could have had potential threats to those he had dealt with. Khulisa worked closely with relevant schools to manage potential related risks as well as to create the necessary awareness.

4.1.2 Victories

The administration of daily data quality checks (both manual and statistical) vastly improved the quality of data collection overall. Having this information on a daily basis meant that problem areas and problem fieldworkers could be identified in near real time and issues addressed accordingly.

4.1.2.1 Completing data collection timeously

We have managed to collect data within the 3 week data collection period.

4.1.2.2 Fieldworker and supervisor recruitment

We have amended our recruitment approach for the 2019 EGRS Data Collection. The decision to shortlist a younger group of fieldworkers proved to be successful.

4.1.2.3 Administration of exam-style fieldworker competency assessment

Led by the DBE, the exam-style fieldworker competency assessment worked well. It also helped the management team to manage and support fieldworkers according to their needs.

4.1.2.4 Working relationship with the DBE

The working relationship between the DBE and Khulisa has contributed to the success of the EGRS Wave 4 2019 Data Collection assignment.

4.1.3 Lessons and Recommendations

From the data quality checks, the following lessons learned are recorded and recommendations:

- Develop and implement structured exam-style fieldworker competency assessments to ensure accurate tracking of fieldworker training progress.
- Although Tangerine is necessary for the reading assessment, it may not be the ideal tool for all assessments, as it cannot be pre-populated with the existing records of the learners to be assessed (which would obviate/minimise the requirement for data matching and merging).
- A data capture platform like REDCap could be considered; it can be pre-populated with all the

learner/school information; assessment/interview data can be added live to the correct record, and the salient outcomes from the reading assessment (from Tangerine) can also be entered into the correct record.

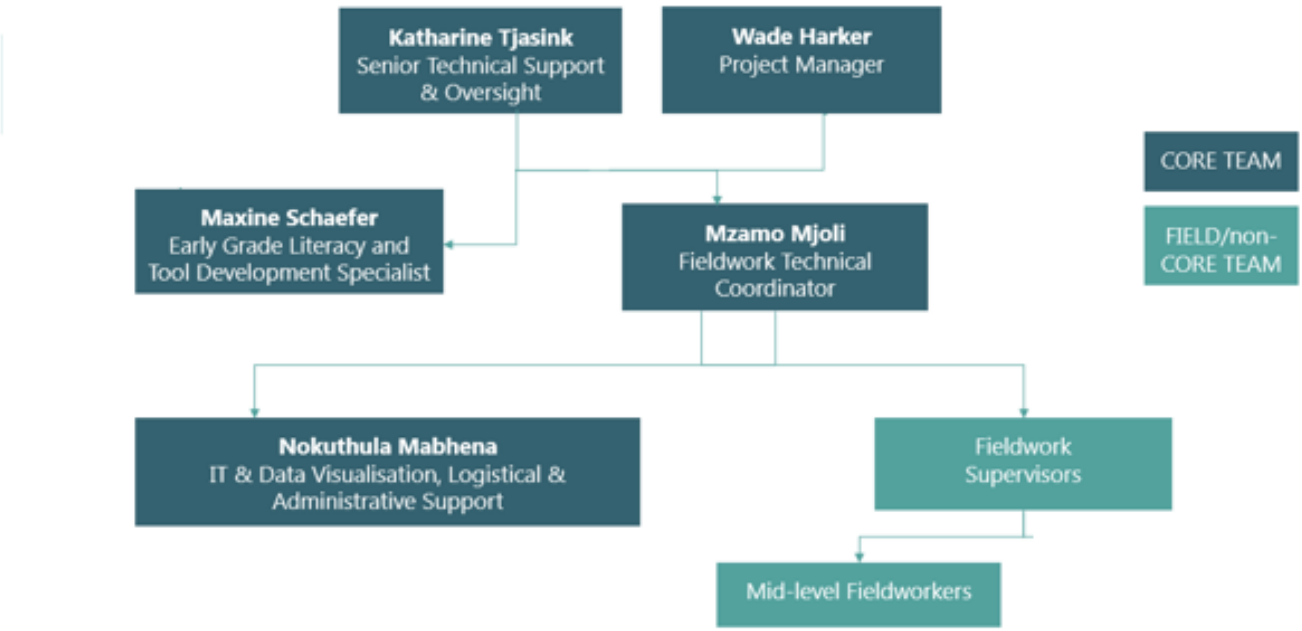
- Although fieldworker training focused heavily on the accuracy and consistency of the capturing of linking IDs, EMIS numbers, school names, teacher names, and learner names, further training could be beneficial.
- If possible, school names should be aligned between the fieldwork schedule, master file and linking file prior to the start of fieldwork.
- If tablets are being updated during data collection, ensure accurate version control and additional checking mechanisms.
- The entry of the linking ID in all instruments should be limited to six alphabetic uppercase characters, to at least prevent the entry of national IDs and linking IDs, which are too long/short.
- If an absent/transferred learner is in fact assessed (elsewhere or at a later date), the linking form must be updated to reflect this; otherwise the information in the linking form is inaccurate and cannot be relied upon for data merging and cleaning.
- The linking file could be used to record whether both the oral and written assessments took place and why exceptions occurred; this could assess in data matching investigations and data cleaning.

5 Project Team

The core Project Team consists of an experienced group of individuals with highly relevant experience to this assignment, including executing data collection for the DBE and USAID.

The team structure is illustrated below in Figure 3, followed by a brief description of each core team members' experience on the following pages.

Figure 3: Core Project Team Organogram



With over 12 years of relevant experience, Katharine Tjasink provided **Technical Oversight and Support** to the project. An Associate Director at Khulisa, Katharine has experience across a range of sectors, including Education, Agriculture, Health, and Youth-development. She has managed research and evaluation assignments at Khulisa since joining on a full-time basis in 2014. This includes the a three-year USAID|Southern Africa Impact Evaluation of the School Capacity and Innovation Program (SCIP) Annual Program Statement recipient award, "*Strengthening Teaching of Early Language and Literacy in Grade R*" (STELLAR).

Additionally, Katharine has overseen assignments across the African continent, which placed emphasis on electronic data collection using smart devices and ODK. She has been involved with the process of assessing and tracking foundation phase learners over various points of data collection, and is aware of the associated challenges.

Mr Wade Harker served as Project Manager for this assignment[4]. Mr Harker specializes in leading large-scale field research assignments in education, agriculture, climate change, and sustainable development sectors. Previously a Senior Associate in the Education and Social Development Division of Khulisa Management Services, Wade has more than 7 years' experience in M&E, evaluations, assessments, and research assignments. He has expertise in managing and overseeing field research across Africa, especially South Africa, managing and supervising large teams of enumerators, and ensuring ongoing data quality.

Mr Harker has overseen multiple large-scale data collection assignments, including the EGRS II data collection assignments in 2017 and 2019. Additionally, he managed the implementation of data collection for the EGRS I data collection in North West in 2018. Mr Harker has overseen the assessments of more than 15 000 learners and he has trained more than 350 researchers to date.

Mzamo Mjoli served as **Fieldwork Coordinator** for this assignment. He is a social scientist and development professional with 8+ years' relevant experience in large-scale education and community engagement projects in South Africa with large-scale data collection, oversight, and training expertise.

Mr Mjoli has a keen understanding of the South African education sector, serving as Data Collector/Evaluator and Fieldwork Supervisor with Khulisa on two USAID-funded ECD projects in the last year, both including data collection in South African primary schools. This includes the recent Midline data collection for EGRS II in Mpumalanga with Khulisa (for the DBE, USAID/SA, and Wits Health Consortium), during which he fulfilled a vital role, and the 2018 EGRS Wave 3 in Mpumalanga, for which he served as fieldwork coordinator. He is fluent in isiZulu and English.

[4] Mr Harker is the Co-owner and Managing Director of Decipher Data (Pty) Ltd, and provides consulting and advisory services to Khulisa on this assignment.

Maxine Schaefer is an **Early Grade Literacy Researcher and Reading Assessment and Tool Development Specialist** with several years' experience in developing and piloting early grade literacy assessments in African languages and English in South Africa. She recently worked as Project Associate on the Second EGRS for Wits Health Consortium for Research Coordination, M&E branch for the DBE, and brings extensive institutional knowledge from EGRS assignments, including the previous EGRS II Mpumalanga data collection assignments.

Ms Schaefer served as a learner assessment tool developer for the EGRS I and II, and developed learner tests for Grade 3: Setswana, English, and Maths items. She also serves as Lecturer of Applied Linguistics at the Department of Linguistics and Modern Languages, at the University of South Africa (UNISA), and has completed several relevant publications. She holds a Master of Arts degree, and is currently enrolled for D Litt et Phil in Linguistics.

Nokuthula Mabhena provided **IT and Data Visualisation** as well as **Logistical & Administrative Support.** Aside from providing IT support relating to mobile data collection software, assisted the PM with data visualisation activities pertaining to project reporting. Nokuthula has worked as a Mobile data collection Software Administrator on various large-scale data collection projects, including impact evaluations for USAID Southern Africa, which required stringent data security and quality protocols.

Apart from the core team members listed above, Khulisa utilized additional consultants, as well as fieldworkers and supervisors, to carry out and quality assure this assignment.

Annex 1. Agreed Data Cleaning Processes

Khulisa needs to show the thinking as to how the data was changed. Changes will be tracked in such a way that any external analyst could replicate the cleaning activity.

Khulisa EGRS II Wave 4 data cleaning responsibility:

- Khulisa to submit 8 cleaned datasets, namely:
    - isiZulu Learner Assessment Dataset
    - isiZulu Learner Written Assessment Dataset
    - SiSwati Learner Assessment Dataset
    - SiSwati Learner Written Assessment
    - School Observation Dataset
    - Principal Questionnaire Dataset
    - Teacher Questionnaire Dataset
    - Learner, Teacher and School Linking Dataset

Additionally, Khulisa will submit the data quality tracking dataset, which were used to track data quality throughout the data collection process.

- Khulisa will clean all school, learner, and teacher identifying information. This includes:
    - Rectifying spelling mistakes of school names across datasets
    - Rectifying School EMIS number across datasets
    - Rectifying learner names and surnames across all datasets.
    - Rectifying learner Unique ID's across all datasets
    - Rectifying teacher names and surnames across all datasets
    - Rectifying principal details where necessary
- Khulisa will use class lists and linking forms to aid the cleaning process.
- Khulisa will insert new variables/comments where changes have been made to the existing data. This will allow the client to easily review the initial information vs the corrected/changed information.
- Khulisa will submit a raw data set that is not touched as a backup. Khulisa to submit two data packs:
        - A raw data pack; and
        - A "cleaned" data pack
- For duplicate entries/submissions, Khulisa will not delete duplicates. Khulisa will highlight potential duplicates.
    - Khulisa to place an extra column where a code 1 and 0 are used. E.g. 1= duplicate and 0= not a duplicate

Annex 2. Data Quality Checks Report

Background

Khulisa was contracted to collect data for this project from 180 schools in two districts in Mpumalanga. Several instruments were administered to learners, teachers, and principals, and raw data sets will be submitted to the DBE upon completion of the fieldwork.

The objective of the data quality aspect of the project was to provide statistical quality assurance services to the DBE, to identify and report any risks to the successful implementation of fieldwork.

Two distinct data quality-checking exercises were performed:

- A series of daily data quality checks was performed on the data from each of the 12 main days of data collection.
- A series of weekly data quality checks performed on the data from each of Weeks 1-3 of data collection.

Daily data quality checks

Process

On each day following a day of main fieldwork (excluding Friday catch-up days), all the learner oral assessment, learner written assessment, teacher interview, principal interview, school observation and linking files were submitted for evaluation, together with the fieldwork schedule and master file of learners to be evaluated received at the start of fieldwork.

Data quality checks were carried out according to a list of checks agreed with the client prior to the start of fieldwork. The checks were programmed in Stata. Results were output to an HTML file and the salient findings were also summarised in an Excel file ("EGRS data checks tracking Day XX"), both of which were sent by the statistician to the Khulisa Technical Representative and the Fieldwork Provider, on the same day.

Findings

The overall findings are summarised alongside each data quality check in Table 1. The detailed findings for each day of fieldwork may be found in the HTML output files and the daily tracking file.

The data checks highlight many inaccuracies in the entering of linking IDs, EMIS numbers and particularly school names and teacher names (also, learner names, although this was not specifically checked). This will make data cleaning unnecessarily time-consuming and tedious; some discrepancies may never be resolved.

Data checks relating to duration, time, and completion of assessments, as well as on outliers in numeric data and skip logic, did not flag any major, consistent concerns. It is therefore assumed that the outcome data is of reasonable quality and will be fit for use once the data matching process is complete.

Table 1: Summary of findings for each daily data quality check

Data quality check no.*	Data quality check	Findings	Approximate % of data affected
Linking file			
-	Number of learners	-	
DC4-a	Number of linking forms per school	Fewer than 15 schools on days 8 and 9	
DC4-b	List any empty fields (subsequently updated to exclude national ID and vocab)	0-12 learners per day listed	<5%
DC4-d	School name: Mismatches vs fieldwork schedule	>100 mismatches each day: Many school names in the two files were not aligned; this is not a consequence of fieldwork as it is internally controlled and could easily have been avoided	>40%
DC4-d	EMIS number: Mismatches vs fieldwork schedule	Mismatches on days 5,8,9,12; updated fieldwork schedule was not provided	
DC4-e	Linking ID: Mismatches vs master list	No mismatches	
DC4-f	Teacher name: Mismatches vs master list	Not done: master list does not contain teacher names	
Learner oral assessment			
-	Number of assessments	-	
DC1	Number of learners assessed per school	Several incorrect EMIS numbers noted on most days (mostly obvious mistyping) Blank records (no EMIS number/linking ID and no other data) were also documented	<5%
DC2	School name: Mismatches vs fieldwork schedule	>100 mismatches each day, ranging from misspelling the name of the school itself to not entering the full name (ABC PRIMARY SCHOOL) of the school correctly	>40%
DC3	EMIS number: Mismatches vs fieldwork schedule	1-12 mistyping errors per day; plus schools visited which deviated from the	<2%

Data quality check no.*	Data quality check	Findings	Approximate % of data affected
		fieldwork schedule	
DC3	School name-EMIS combination vs fieldwork schedule	>100 mismatches each day resulting directly from the above mismatches	>40%
DC4-e	Linking ID: Mismatches vs master list	4-16 mismatches per day, resulting from entries other than the linking ID (e.g. names, national ID numbers) and mistyping of the linking ID	<5%
DC4-f	Linking ID: Mismatches vs linking file	65-101 learners per day in linking file but not in oral assessment (combination of unavailable learners and non-matching linking IDs); 4-26 learners per day in oral assessment but not in linking file (investigated in weekly data checks)	20-30% <10%
DC4-g	Teacher name: Mismatches vs linking file (based on linking ID match)	On top of the linking ID mismatches, there were many teacher name mismatches, ranging from mistyping errors, adding second names, switching name/surname and then gross mismatches	
DC8-a	Duplicate learners (based on linking ID)	1-4 duplicates per day; some were blank records, some not (investigated in more detail in the weekly data checks)	<2%
DC8-b	Assessments outside school time	0-2 assessments before/after school time each day (but close to school start/end)	<1%
DC8-c	Completion of assessment	0-3 incomplete assessments per day	<1%
DC8-d	Total duration of assessment (Task 1-8): listing of assessments below 5th percentile and above 95th percentile	Listed	
DC8-e	Words per minutes (Tasks 1-5): listing of assessments below 5th percentile and above 95th percentile	Listed; some high values for Task 4 observed on some days	
DC8-f	High levels of missing observations	0-2 assessments with non-zero number of missing observations per day	<1%

Data quality check no.*	Data quality check	Findings	Approximate % of data affected
Learner written assessment			
-	Number of assessments	Usually fewer than oral assessments (investigated in more detail in weekly data checks)	
DC1	Number of learners assessed per school	Several incorrect EMIS numbers noted on most days (mostly obvious mistyping) Fewer than 15 schools (excluding obvious mistyping of EMIS numbers) found on approximately half the days; presumably capturing of these assessments was done later Blank records (no EMIS number/linking ID and no other data) were also documented	<5%
DC2	School name: Mismatches vs fieldwork schedule	>100 mismatches each day, ranging from misspelling the name of the school itself to not entering the full name (ABC PRIMARY SCHOOL) of the school correctly	>40%
DC3	EMIS number: Mismatches vs fieldwork schedule	1-3 mistyping errors per day; plus schools visited which deviated from the fieldwork schedule	<2%
DC3	School name-EMIS combination vs fieldwork schedule	>100 mismatches each day resulting directly from the above mismatches	>40%
DC4-e	Linking ID: Mismatches vs master list	5-12 mismatches per day, resulting from entries other than the linking ID (e.g. names, national ID numbers) and mistyping of the linking ID	<5%
DC6	Linking ID: Mismatches vs oral assessment	16-35 records per day in written , not in oral, assessment 18-39 records per day in oral, not in written, assessment (investigated in more detail in weekly data checks)	5-20% 5-20%
DC6	EMIS number: Mismatches vs oral assessment	1-5 mismatches per day (plus those which did not match on linking ID)	<2%
DC6	Teacher name: Mismatches vs oral assessment	Not done: written assessment does not contain teacher names	

Data quality check no.*	Data quality check	Findings	Approximate % of data affected
DC9-a	Duplicate learners (based on linking ID)	1-5 duplicates per day	<2%
DC9-b	Completion of assessment	0-6 incomplete assessments per day	<2%
DC9-c	High levels of missing observations	0-6 assessments with non-zero number of missing observations per day	<2%
Teacher interview			
-	Number of interviews	-	
DC1	Number of teachers interviewed per school	More or fewer than 15 schools found on approximately half the days Blank records (no EMIS number/teacher name or any other data) were also documented	
DC2	School name: Mismatches vs fieldwork schedule	15-22 mismatches each day, ranging from misspelling the name of the school itself to not entering the full name (ABC PRIMARY SCHOOL) of the school correctly	>50%
DC3	EMIS number: Mismatches vs fieldwork schedule	0-2 mistyping errors per day; plus schools visited which deviated from the fieldwork schedule	<10%
DC3	School name-EMIS combination vs fieldwork schedule	15-22 mismatches each day resulting directly from the above mismatches	>50%
DC4-f	Teacher name: Mismatches vs linking file	Many teacher name mismatches, ranging from mistyping errors, adding second names, switching name/surname and then gross mismatches	
DC10-a	Duplicate teachers (based on teacher name + surname)	0-2 duplicate teachers each day	<10%
DC10-b	Interviews outside school time (up to 14:30)	0-1 interviews before/after school time each day (but close to school start/end)	<5%
DC10-c	Completion of interview	0-4 incomplete interviews each day	<15%
DC10-d	Exceptions to skip logic	Q4.4.1 to Q4.4.2: 0-1 exceptions each day Q5.2 to Q5.3.x: 0-5 exceptions each day	<5%

Data quality check no.*	Data quality check	Findings	Approximate % of data affected
			<20%
DC10-e	Numeric data: listing of interviews with data below 5th percentile and above 95th percentile	Q9.3.1-3 had very high values on some occasions	<5%
DC10-f	High levels of missing observations	0-2 interviews with non-zero number of missing observations per day	<10%
Principal interview			
-	Number of interviews	-	
DC1	Number of principals interviewed per school	Fewer than 15 schools found on approximately half the days Blank records (no EMIS number/principal name or any other data) were also documented	
DC2	School name: Mismatches vs fieldwork schedule	8-15 mismatches each day, ranging from misspelling the name of the school itself to not entering the full name (ABC PRIMARY SCHOOL) of the school correctly	>50%
DC3	EMIS number: Mismatches vs fieldwork schedule	0-1 mistyping errors per day; plus schools visited which deviated from the fieldwork schedule	<10%
DC3	School name-EMIS combination vs fieldwork schedule	8-15 mismatches each day resulting directly from the above mismatches	>50%
DC11-a	Interviews outside school time (up to 14:30)	0-1 interviews before school time each day (some before 07:00)	<10%
DC11-b	Completion of interview	0-1 incomplete interviews each day	<10%
DC11-c	Exceptions to skip logic	Exceptions were listed Q5.4.2 and Q5.4.3 are allowing text entries	
DC11-d	High levels of missing observations	None	

Data quality check no.*	Data quality check	Findings	Approximate % of data affected
School observation			
-	Number of observations	-	
DC1	Number of observations conducted per school	Fewer than 15 schools found on approximately half the days Blank records (no EMIS number/school name or any other data) were also documented	
DC2	School name: Mismatches vs fieldwork schedule	7-11 mismatches each day, ranging from misspelling the name of the school itself to not entering the full name (ABC PRIMARY SCHOOL) of the school correctly	>50%
DC3	EMIS number: Mismatches vs fieldwork schedule	0-1 mistyping errors per day; plus schools visited which deviated from the fieldwork schedule	<10%
DC3	School name-EMIS combination vs fieldwork schedule	7-11 mismatches each day resulting directly from the above mismatches	>50%
DC11-b	Completion of observation	0-1 incomplete observations each day	<10%
DC11-d	High levels of missing observations	None	

** As referred to in agreed list of checks, and in HTML output*

Weekly data quality checks

Process

Each week following a full week of fieldwork, the following investigations were done using all data collected in the linking file, learner oral assessment, and learner written assessment from the Monday to Sunday of the previous week:

- The number of school and number of learners to be assessed was determined from the linking file.
- Using the "Comments" field in the linking file, learners who were marked absent, ill, transferred, or untraceable were removed from consideration.
- The learner oral assessment data file was stripped of records, which were empty, started before 07:30, and were duplicate linking IDs with missing data. Remaining duplicates (by linking ID) were documented.
- Next, the learner oral assessment file and linking file were merged by linking ID.
    - Mismatches due to linking ID mistyping were corrected by hand as far as possible.
    - The number of learners in the linking form, but with no oral assessment, was documented.
    - The of learners with oral assessments who were not in the linking file, was documented, and reasons for the existence of these records were sought.
    - Next, among the records that matched directly on linking ID, the number of true teacher name mismatches (excluding spelling errors, inclusion of second name, etc.) was determined.
- The learner written assessment data file was stripped of records that were empty, and were duplicate linking IDs with missing data. Remaining duplicates (by linking ID) were documented.
- Next, the learner oral and written assessment files were merged by linking ID.
    - Mismatches due to linking ID mistyping were corrected by hand as far as possible.
    - The number of learners in the linking form, but with no oral assessment, was documented.
    - The of learners with written assessments who were not in the oral assessment file, was documented, and reasons for the existence of these records were sought.

The checks were programmed in Stata. The findings were summarised in an Excel file ("EGRS investigations Week X"), which was sent by the statistician to the Khulisa Technical Representative and the Fieldwork Provider.

Findings

The overall findings are summarised for each week in Table 2. The detailed findings for may be found in the weekly investigation file.

The data checks highlight that approximately 10% of the assessable learners did not have an oral assessment; the reasons for this should become clearer once the data set is cleaned in its entirety. Conversely, the primary reason learners appeared to have oral assessments but were not found in the linking file was that they were marked absent/transferred in the linking file and thus did not appear to qualify for assessment. True teacher name mismatches between the oral assessments and the linking file appeared to decrease over the course of fieldwork.

Similarly, there are learners who had only one of the oral and written assessments, but not both; the reasons for this should become clearer once the data set is cleaned in its entirety. It does not appear, however, that learners not in the study were assessed.

Table 2: Summary of findings for weekly investigations

Investigation	Week 1	Week 2	Week 3
Linking file			
Number of schools to be assessed	59	59	61
Total number of learners to be assessed	1094	1101	1119
Number of learners to be assessed after removal of absent, ill, transferred, untraceable learners	910	927	913
Learner oral assessment			
Raw data file	926	898	881
After removal of blanks, assessments before 07:30, and duplicates (by linking ID)with missing data	884	859	844
Duplicates remaining (according to linking ID)	10	4	9
Linking ID matches with linking file (either direct matches of linking ID or mistyping matched by hand)	853	851	833
Learners in linking file but not in oral assessment	68 (8.0%)	82 (9.6%)	90 (10.8%)
Learners in oral assessment but not in linking file	31	8	11
School missing from linking form	*13*		
Learner marked absent/transferred in linking form	*9*	*5*	*9*
No match for learner in linking form	*3*		
School not listed in linking form	*4*		
Test data	*2*		
School not assessed in this week		*1*	*2*
Muddled records		*2*	
Teacher name mismatches with linking file (out of records with direct linking ID match)	13.8%	5.4%	3.8%

Investigation	Week 1	Week 2	Week 3
Learner written assessment			
Raw data file	875	860	848
After removal of blanks, assessments before 07:30, and duplicates (by linking ID)with missing data	866	854	842
Duplicates remaining (according to linking ID)	6	5	8
Linking ID matches with oral assessment (either direct matches of linking ID or mistyping matched by hand)	843	818	815
Learners in oral, but not written, assessment file	48	46	35
Learners in written, but not oral, assessment file	29	41	36
Learner is in linking file	*17*	*36*	*36*
Learner is NOT in linking file	*8*		
Muddled records	*1*		
School not listed in linking file	*1*		
Test data	*2*		
School not assessed in this week		*4*	
Learner is in linking file but listed as transferred		*1*	

Annex 3. Data Quality Assessment and Surveillance Plan (QASP)

In addition to the data quality assessment measures outlined in this report, the following quality procedures were observed:

1	PROJECT MANAGEMENT – GENERAL QUALITY STANDARDS	
1.1	**Documentation**	a. Version control is used consistently for all documentation.
		b. All project documentation (reports, presentations, etc.) undergo internal quality reviews by Khulisa with a focus on • ensuring responsiveness to client requirements, and • high quality writing (e.g. coherence, clarity, well developed ideas, internal logic, use of strong sentence construction with active voice and smooth transitions, correct spelling/grammar/punctuation, good formatting, visualizations etc.)
		c. Feedback/comments from external reviewers (and Khulisa subsequent actions on these) are documented and responded to in a timely manner
		d. All information generated by the project is accurate, reliable, clear, complete, unbiased, and useful. Every phase of information development (creation, collection, maintenance, and dissemination) is governed by these information quality principles.
1.2	**Meetings (face-to face and/or telephonic)**	a. An inception meeting is held to ensure the needs of the client are understood
		b. Ongoing telephonic and face-to-face meetings are carried out throughout the assignment to ensure the client is aware of risks, mitigation plans, issues confronted in the field, data quality issues, etc.
1.3	**Work plan**	a. Work plans include: key activities within each phase of the assignment, project time lines/Gantt charts, and responsibilities.
		b. Work plans are reviewed regularly and deviations are flagged, explained, and addressed in consultation with relevant staff and stakeholders.

PHASE 1: INCEPTION		
2	**STAFFING AND TRAINING**	
2.1	**Staffing**	a. Staff recruitment is done in a competitive and transparent manner, to ensure the most appropriate candidates are selected.
		b. Data collectors are chosen based on the following key criteria: • Ability to read and speak the language of the assessment • Previous experience administering assessments and collecting data • Proficiency using a smart phone or tablet • Ability to follow written instructions and read maps • Other criteria, as specified by the client
		c. Field supervisors are recruited based on the following key criteria: • Prior work experience with fieldwork supervision and/or leadership • Organizational ability and attention to detail • Strong computer skills including ability to download and review electronic data • Experience with research and data collection - including quality control • Other criteria, as specified by the client
		d. A job description with clear job requirements is in place for all positions.
		e. Penalties are built into all staff contracts to ensure timeliness and completeness as per the established due dates of the approved work plans.

		f. More data collectors and fieldworker supervisors are recruited than needed to ensure that the most competent individuals are deployed and to allow for a list of alternates in case individuals drop out.
		g. A data base of all enumerators is maintained to keep track of their performance (timeliness and quality) and only those who have a record good performance will be engaged in subsequent assignments.
2.2	**Training**	a. All training activities have clearly articulated training objectives and agendas.
		b. The length, methods, and content of trainings are relevant to and sufficient for the research objectives and reflect best practices.
		c. Training of data collectors and supervisors includes • the use of the data collection tools • entering data into the electronic devices • quality assurance, and best practices for conducting surveys
		d. Training of data collectors includes guidelines on how to physically protect data collection devices, including e.g. confer liability of the device to data collectors, take precautions when transporting and storing devices.
PHASE 2: IMPLEMENTATION		
3	**DATA COLLECTION**	
3.1	**Fieldwork Logistics**	a. Fieldwork schedules include realistic timelines for on-site data collection as well as travel between sites
		b. Officials at selected sites are notified of the fieldwork visit sufficiently in advance
		c. Accommodation and transportation arrangements are confirmed prior to fieldwork commencement
		d. Clear communication protocols are shared with the fieldwork team prior to deployment

3.2	**Fieldwork Ethics**	a. Consent (and its recording) is obtained without participant coercion
		b. Each fieldworker signs non-disclosure agreements as well as a declaration of conflict of interest
3.3	**Fieldwork Monitoring/ Oversight**	a. Fieldwork Supervisors observe data collectors as they conduct interviews and other measures, noting errors and misconceptions and taking remedial action where necessary. Fieldwork supervisors discuss the identified errors with data collectors at the end of every data collection day to correct identified issues and errors
		b. Fieldwork Supervisors check completion of submitted tools and identify and resolve any data quality issues
3.4	**Data Quality and Security**	a. The quality of collected data is ensured by: • Daily checks (statistical and manual) checking for such parameters as completeness, timeliness, matching of IDs, etc. • Sign-off from Supervisors which indicate that the data was checked for quality.
		b. Data Security is ensured by: • Using only approved devices for data collection. • Protecting devices with user login and passwords. • Backing up collected data on secure Khulisa servers. • Keeping paper-based data in protected boxes
4	**DATA CLEANING AND ANALYSIS**	
4.1	**Data Consolidation and Cleaning**	a. The original data set(s) is downloaded to the secure Khulisa server and saved as protected "DO NOT TOUCH - Protected Master (Raw)" file

		b. A copy of the original data set(s) is cleaned using the following procedures: • Rectifying spelling mistakes of school names across datasets • Rectifying School EMIS number across datasets • Rectifying learner names and surnames across all datasets. • Rectifying learner Unique ID's across all datasets • Rectifying teacher names and surnames across all datasets • Rectifying principal details where necessary
		c. The final cleaned data set is saved on the secure Khulisa server as a "DO NOT TOUCH - Protected Master (Cleaned)" version.
		d. Data set(s) version control is implemented
5	**General**	
5.1	**General**	a. All reports undergo internal quality reviews by Khulisa with a focus on • ensuring responsiveness to client requirements, and • high quality writing (e.g. coherence, clarity, well developed ideas, internal logic, use of strong sentence construction with active voice and smooth transitions, correct spelling/grammar/punctuation, good formatting, visualizations etc.).
		b. The reports are presented to all relevant stakeholders in a clear and accessible manner
		c. Feedback from stakeholders and reviewers (and Khulisa actions) are documented and responded to in a timely manner