GCRO Quality of Life Survey V (2017/18): Data Report

07 June 2019

1 Introduction

This report serves as a summary of changes, amendments, recodes and corrections made during the data validation process for the fifth Quality of Life (QoL V) survey (2017/18) implemented by the Gauteng City-Region Observatory (GCRO).

The report should be reviewed in conjunction with the completed SPSS dataset, the questionnaire, and the field report.

2 Data collection background

2.1 Research instrument

The research instrument was designed by the GCRO, with input from a wide range of stakeholders. The final questionnaire included 248 questions, divided into 15 sections. All questions were close-ended. The 15 sections of the questionnaire included:

1) Dwelling and household information, and access to services;
2) Satisfaction with services;
3) Migration;
4) Neighbourhood or community;
5) Transport;
6) Internet access and use;
7) Household characteristics;
8) Public participation and satisfaction with government;
9) Social and political views and opinions;
10) Satisfaction with life;
11) Business and employment;
12) Crime and safety;

13) Community participation and protest;

14) Health;

15) Further demographic, household, and personal information.

The questionnaire was developed in English, and key concepts were translated into isiZulu, isiXhosa, Setswana, Sesotho and Afrikaans by the linguistics department at the University of Johannesburg (UJ). The instrument and translations were workshopped between ResearchGO and GCRO, to finalize the language used and ensure a clear, shared understanding of all questions prior to training. Fieldworkers were trained on translations, and the translations were available to them on their data collection device and in hard copy. Please refer to the Field Report for further detail on these processes.

2.2 Data collection system

The survey was programmed in the ResearchGo platform. Data collection was implemented using the ResearchGo application on tablet devices. The ResearchGo platform and application were developed by Relentless Technologies, specifically for data collection purposes. The platform required significant upgrades in order to cope with the skip patterns and question types required for QoL.

The ResearchGo platform supported the full data collection process, from initial navigation to a selected survey point, through to selection of the appropriate respondent and implementation of the survey itself. Where multiple visits to the same survey point were required, this was also supported by the system. Within the system, questions were implemented as a series of pages, which contained a small number of questions. For each of these pages, the platform automatically recorded duration, as well GPS coordinates, and any unusual behaviours such as changes to response options.

Fieldworkers were able to complete questionnaires regardless of network coverage. GPS coordinates were satellite based, and were consequently independent of network coverage.

2.2.1 Question types

The default question type on the ResearchGo platform provides the interviewer with the question text, and a list of response options from which to select a single option. This was used for the majority of questions.

Two additional question types were used for questions in which respondents were able to select multiple responses. Firstly, for questions allowing multiple responses, but not requiring the fieldworker to read all options out to the respondent, the default format was used with a slight

modification to allow selection of multiple response options. In the questionnaire, these questions are specified by “Coding note: multiple mention”. See for example Q5.6, where respondents were asked to list all modes of transport used in their most frequent trip.

Secondly, for questions in which each of the multiple response options needed to be read out to the respondent, so that they could indicate whether or not that option applied, a ‘Yes/No’ format was used. This provided the data interviewer with the question text, followed by the list of response options, with a ‘yes’ and a ‘no’ button next to it. The questionnaire would not proceed until either the yes or no button was selected for each response option. In the questionnaire, these questions are specified by “Coding note: Yes/No list”. An example of this type of question is Q4.7, in which respondents indicated whether or not they lived within a 15 minute walk of various services.

Q5.3 (destination for most frequent trip) was particularly challenging to implement. Further details are provided in Section 2.3, below.

Range limits were placed on most numerical input questions, to reduce the chance of inadvertent entry of an inaccurate response. Limits are specified in the questionnaire whenever they were implemented. Only two numerical questions (Q1.22 and Q1.24 – the number of bags of refuse and recycling generated by the household) allowed for the entry of decimal numbers, while all other numerical questions accepted only integer values.

All questions were implemented to require a response unless intentionally skipped through survey design, meaning that questions could not accidentally be left blank. As documented in Section 3 below, questions which were intentionally skipped through survey design were coded with ‘-1’. For potentially sensitive questions which a respondent might not want to answer (see in particular Q15.21 – household income – and Q15.22 – political preferences), a “respondent refused” option was available for selection when a respondent did not wish to respond.

2.2.2 Skip patterns and logic checks

Skip patterns were implemented to ensure that, in so far as possible, inapplicable questions were not asked of respondents. Unfortunately, due to limitations of the data collection platform, it was not possible to implement all skip patterns as planned. Consequently, in certain instances, fieldworkers were provided with response options to indicate that a question should have been skipped. This type of option was available in Q1.14, Q1.15, Q1.16, Q1.26, Q1.28, and Q1.29.

It was also not possible to conduct logic checks and generate inconsistency alerts to fieldworkers on a live basis when responses within the questionnaire were inconsistent. Consequently, consistency checks were implemented on questionnaires as they were downloaded from the server, and were queried with fieldworkers at this point. Surveys with serious or extensive internal inconsistencies were not accepted.

2.3 Implementation of Q5.3 (travel destination)

As indicated in Section 2.2.1, implementation of Q5.3 (travel destination of most frequent trip) was particularly complex, and involved a 3-stage process. A respondent was first asked the province of the destination, using the standard question format. A response option for a trip with a destination outside of South Africa was also provided. Respondents who indicated that their destination was in Gauteng, Free State, Limpopo, Mpumalanga or North West were subsequently asked to identify the Municipality in which their destination was located. Those who had selected other provinces or another country were not asked for further information. In those instances in which the destination Municipality was requested, predictive text and a drop-down menu of municipalities in the selected province were used. As the fieldworker began to type the municipality name, the available options would be limited. For Gauteng provinces, the five Metropolitan and District municipalities were included, but Local municipalities were not listed. For other provinces, only selected Metropolitan and Local municipalities were included, together with an 'Other' option. Municipalities that were available for selection are listed in Table 1 below.

Table 1: Municipalities available for selection

Province	Municipalities available for selection
Gauteng	Ekurhuleni (Metropolitan) Johannesburg (Metropolitan) Tshwane (Metroplitan) Sedibeng (District) West Rand (District)
Free State	Mangaung (Metropolitan) Masilonyana (Local) Matjhabeng (Local) Nala (Local) Other
Limpopo	Bela-Bela (Local) Mogalakwena (Local) Polokwane (Local) Thabazimbi (Local) Other
Mpumalanga	Dr JS Moroka (Local) Emalahleni (Local) Mbombela (Local) Thembisile (Local) Victor Khanye (Local) Other
North West	Kgetlengrivier (Local) City of Matlosana (Local) Mahikeng (Local) Other

Once respondents had selected the appropriate municipality, they were asked to provide the sub-place of their destination. Participants who selected ‘Other’ for municipality were not asked to provide any further information. Again, a predictive text drop-down menu was used, comprised of all the main-place and sub-place combinations in that particular municipality.

Experience with previous iterations of the survey has shown that some sub-place names are not intuitive, and confuse respondents. When sub-places were not plausible travel destinations, they were removed from the list of available response options. This was particularly the case with sub-places with the NU suffix, which denotes ‘non-urban’ (see Table 2). So while ‘City of Johannesburg’ would typically be understood as the centre of Johannesburg, ‘City of Johannesburg NU’ refers to a fundamentally different outlying area, unlikely to be the destination for travel. Where a number of sub-places shared very similar names, only the most geographically central was retained to minimise confusion (see Table 3). When sub-place names were confusing, these were edited for clarity (see Table 4 and 5).

Table 2: Sub-places in Gauteng including 'NU' that were removed from the dropdown for Question 5.3

Sub-place code	Sub-place name
798002003	City of Johannesburg NU
797002003	Ekurhuleni NU
760006002	Emfuleni NU
762004002	Lesedi NU
766002002	Merafong City NU
761002002	Midvaal NU
763001002	Mogale City NU
764003001	Randfontein NU
799026001	Tshwane NU
765004002	Westonaria NU

Table 3: Sub-places including 'SP' that were removed from the dropdown of Q5.3

Sub-place code	Sub-place name
799059090	Centurion SP1
799059089	Centurion SP2
799059002	Centurion SP3
799059012	Centurion SP4
799035104	Pretoria SP
764002031	Randfontein SP1
760009006	Vereeniging SP1
760009026	Vereeniging SP2

Table 4: Sub-places in Gauteng that were renamed to include 'Central/CBD' in the dropdown of Q5.3

Sub-place code	Sub-place name	Sub-place name used in drop down list

798015089	Johannesburg SP	Johannesburg central/CBD
797006020	Kempton Park SP	Kempton Park Central/CBD
760009032	Vereeniging Central	Vereeniging Central/CBD
762014004	Heidelberg Central	Heidelberg Central/ CBD
766004009	Carletonville Central	Carletonville Central/CBD
761006012	Meyerton Central	Meyerton Central/CBD
763004038	Krugersdorp Central	Krugersdorp Central/CBD
764002017	Randfontein SP	Randfontein central/CBD
799035058	Pretoria Central	Pretoria Central/CBD
765008001	Westonaria SP 1	Westonaria Central/CBD
797007010	Edenvale SP	Edenvale Central/CBD
797026002	Tsakane SP	Tsakane Central/CBD

Table 5: Sub-places outside Gauteng that were renamed to include 'mines & farms' in the dropdown of Q5.3

Sub-place code	Sub-place name	Sub-place name used in drop down list
499002001	Mangauang NU	Mangauang - mines & farms
667002001	Mafikeng NU	Mafikeng - mines & farms
868010002	eMalahleni NU	eMalahleni - mines & farms
874003001	Mbombela NU	Mbombela - mines & farms
974002003	Polokwane NU	Polokwane - mines & farms
977001001	Thabazimbi NU	Thabazimbi - mines & farms
981002002	Bela-Bela NU	Bela-Bela - mines & farms

The data collection platform internally concatenated all responses to the 3 sub-questions in Q5.3, and we received this single variable in the raw dataset. This raw data is retained in the variable ‘Q5.03_destination’. To simplify use, we have created two additional variables for each of the four levels of geography (province, municipality, main-place and sub-place). The first variable for each level of geography provides a text string with the name of the area. These strings are the names as they were included in the data collection application, and deviate from the official names as details in Tables 2-5 above. The second variable for each level of geography is the official numeric area code, and is labelled with the official area name. When a participant was not asked about a particular level of geography, both relevant variables are coded ‘-1’ (the standard indicator for an intentionally skipped question, as per Section 3 below).

The approach to data collection for this question generally worked well, but proved problematic for individuals travelling to Duncanville, in Gauteng. This area has a consistently named main-place (Vereeniging) and sub-place (Duncanville) in each of two different local municipalities (Emfuleni and Midvaal) within the Sedibeng District municipality. As local municipality information was not collected, it was not possible to determine to which local municipality, and by extension the sub-place and main-place, the respondent was referring. For these cases, we have left the area names, as collected, in the text versions of the main-place and sub-place questions, but have used ‘-3’ (denoting data missing due to a fieldwork/system error) for the main-place and sub-place codes, as well as the local municipality. This affects 25 surveys.

2.4 Interview length

On average, an interview took 40.5 minutes. This duration excludes respondent selection processes, and the questions asked of fieldworkers at the end of the survey. Figure 1 presents the distribution of interview length. All interviews under 30 minutes in duration were subject to additional manual checks before approval. The final dataset includes 5193 interviews (20.9%) under 30 minutes, of which 56 were between 19 and 20 minutes. No interviews shorter than 19 minutes were accepted. Only 322 interviews (1%) were recorded at more than 90 minutes in length. These were largely due to instances in which the survey was interrupted and subsequently resumed, but also included some surveys where fieldworkers reported extremely talkative or slow respondents.

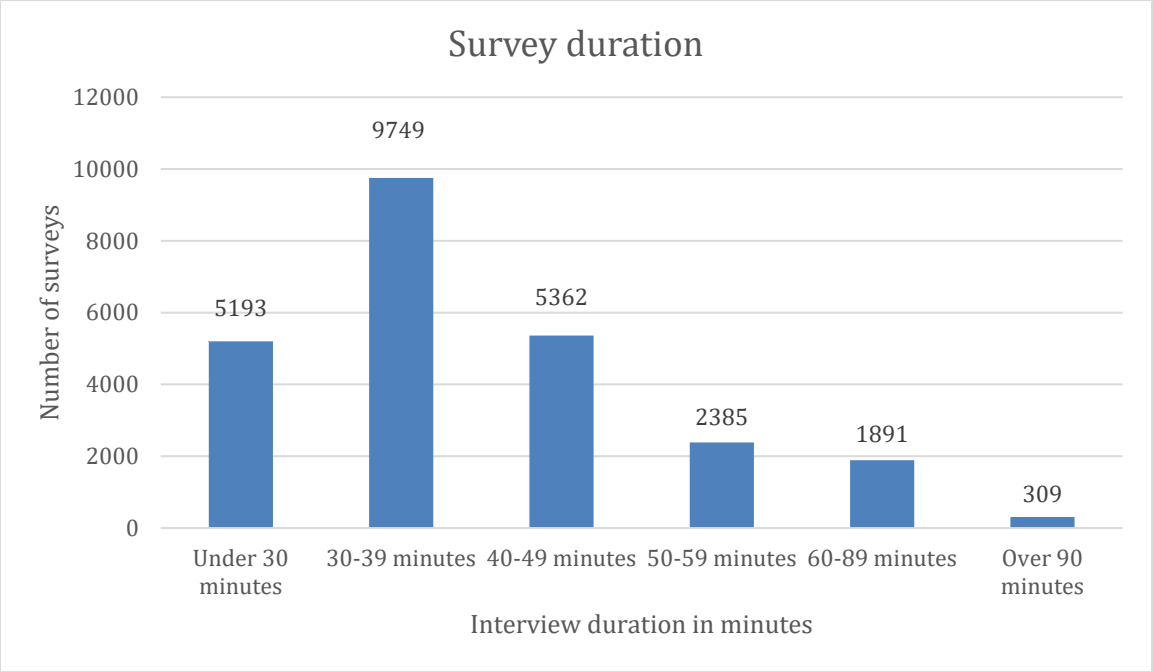

Figure 1: Duration in minutes of surveys in final dataset

2.5 Other specify

In previous iterations of QoL, respondents who selected an ‘other’ response were, in some number of instances, asked to specify. This enabled a proportion of ‘other’ responses to those questions to be recoded into one of the pre-specified responses. This was implemented particularly extensively in QoL IV. In QoL V respondents were not asked in any instances to provide further information when they selected ‘other’, due to the burden of recoding. This does mean that in some instances, the proportion of ‘other’ reflected in the QoL V dataset is higher than in previous iterations of the survey, and in particular, in comparison with QoL IV. The analyst is encouraged to critically consider the proportion of ‘other’ responses for a variable. For

longitudinal analysis, it is important to consider whether ‘other’ responses have been treated equivalently over time.

3 Universal codes

All non-numerical variables in the dataset have been coded numerically. The numerical codes are labelled in the SPSS version of the dataset, and the coding is also documented in the questionnaire itself.

3.1 Missing information

Standard codes are used in the dataset to represent information that is missing for various reasons. These are detailed in Table 6 below.

Table 6: Standard codes for missing data

-1	Data missing due to a valid skip pattern (i.e. the question was not asked of the respondent as it was not applicable to that respondent)
-3	Data missing due to a fieldwork/system error (i.e. the question was not asked of the respondent, but should have been, or a response was not appropriately recorded)

As described in Section 2.3 above, the main question affected by fieldwork/system errors is Q5.3 – destination of most frequent trip (n=25).

3.2 Standard response options

Default coding for questions with standard response options are listed below. However the user should in all instances be guided firstly by the codes provided in the questionnaire and the labelling within the dataset.

Table 7: Default coding for Yes/No questions

0	No
1	Yes

Table 8: Default coding for satisfaction scale questions

1	Very satisfied
2	Satisfied
3	Neither satisfied nor dissatisfied
4	Dissatisfied
5	Very dissatisfied

Table 9: Default coding for agreement scale questions

1	Strongly agree
2	Agree
3	Neither agree nor disagree
4	Disagree
5	Strongly disagree

4 Unique identifier

A unique identifier was automatically generated by the data collection system for every attempted interview, and these codes were retained as the unique respondent identifiers in the final dataset. This code is a primary key in the SPSS dataset. The variable name is resp_id.

5 Implementation challenges related to specific questions

During data collection, it became apparent that a few questions did not work well in the field. Details of these questions, as well as our recommendations for the use of the associated data, are provided below. In general, our approach has been to minimise any adjustments to the dataset, and to always retain the original variable in addition to any recoded variables available. This is to enable each analyst to make a fully informed decision around whether and how to use a variable which may be less accurate than ideal.

Q1.01.01 – number of households in dwelling: both fieldworkers and respondents struggled to differentiate between the number of households in the dwelling, and the number of people living in the household, and consistently provided the number of people for this question. Despite extensive re-training, this confusion could not be remedied. Due to the extent of problematic responses to this question, the variable has been removed from the dataset.

Q1.01.02 – number of people in household: in some instances, fieldworkers only recorded the number of adults in the household, and excluded children from this figure. This was largely remedied through re-training, but data challenges remain for some earlier surveys.

Q1.14, Q1.15 and Q1.16 – additional water sources: as mentioned in Section 2.2.2, it was not possible to skip these questions for respondents who had already listed the water source in question as their main water source in Q1.07. Consequently, an option of “already mentioned as main water source” was provided for these respondents. Unfortunately this option did not work well in the field, and efforts to remedy were not very effective. In many instances, fieldworkers reported that they had understood the option to mean that the respondent had already mentioned their main and only water source, and that this question was consequently not

applicable. In these cases, the already mentioned option was typically selected for all three questions. In other instances, the already mentioned option was used to indicate a positive response for a particular water source, because the fieldworker had understood the option to mean that while the respondent had already identified a main water source, this was an additional one. In these instances, the water source in question was typically not listed in Q1.07. We have provided recodes for these variables, as detailed in Section 6 below. We advise that the recodes be used when possible, with due caution, and that use of the original variables is avoided.

Q1.25 – type of electricity supply: fieldworkers experienced difficulty in differentiating between different types of metered electricity connections, despite extensive training. Furthermore, it was not possible to restrict particular combinations of responses. This means that in a number of instances multiple forms of metered electrical connections were selected. In some cases, fieldworkers reported that they selected all potentially applicable options when they were unsure which to choose. In other instances, one or more options was selected, together with the 'don't know' option, to indicate their lack of certainty. In other instances, fieldworkers reported that respondents used 'don't know' when they were aware that they were using an illegal electricity connection. Finally, there were some reports of instances in which 'connection from elsewhere' was used to indicate that the respondent did not have any electricity connection, and obtained energy in a different way. We suggest that when this variable is used, the analyst may want to combine the various types of metered connections into a single category, or consider using the recode detailed in Section 6 below.

Q1.26, Q1.28 and Q1.29 – electricity supplier, electricity expenditure and electricity interruptions: as it was not possible to skip this question for individuals who reported no access to electricity in Q1.25, a 'not applicable' option was included for use by the fieldworker in these cases. While this option was used correctly in some instances, this was not always the case:

- In some instances, 'not applicable' was used for individuals with electricity. Feedback from the fieldwork team suggests that in some instances is was used as a proxy for 'don't know', when the respondent was unsure of the response. For Q1.26 (electricity supplier), there were some reports that people using pre-paid electricity wanted to provide the name of the shop where they bought electricity, as opposed to one of the available options, and that the not applicable option was used in these cases. For Q1.28 (electricity expenditure) there were some reports that people were reluctant to respond, and not applicable was used in these cases. In addition, some fieldworkers reported using this option when the respondent did not have a formal connection, or were using electricity supplied by other individuals.

- In some instances, people reporting no access to electricity did not make use of the 'not applicable option'. Fieldworkers report that the Q1.28 and Q1.29 were sometimes understood to refer to all energy sources, and not just electricity, meaning that responses were collected even from individuals without electricity. Fieldworkers also reported some individuals who did not report having electricity answered these questions as they were in fact using an illegal electricity connection, but had been reluctant to report this.

We have not provided recodes of these questions, but suggest that analysts consider the concerns detailed above before deciding on the most appropriate way to use this data for any particular analysis.

Q1.27 – main source of lighting: there are 408 surveys which report electricity as the main source of lighting, although the respondent has also indicated that they do not have electricity. The majority of these respondents have also responded to Q1.26 and Q1.28, which suggests they do have access to electricity. One possible explanation for this discrepancy is that a number of individuals with illegal electricity connections may have had difficulty responding to Q1.25 as they were unsure which option to select, or may not have been comfortable disclosing the nature of their connection (see notes regarding Q1.25 above). We have not provided any recodes for this variable, but advise the analyst to consider how best to use this particular variable for any particular analysis.

Q5.06 and Q5.07 – modes of travel for most frequent trip: there are a number of instances in which the travel mode selected in Q5.07 (mode used for longest distance within trip) is not listed in the group of modes selected in Q5.06 (all modes of travel used for trip). A potential explanation for this provided by the fieldwork team is that some fieldworkers and respondents did not understand that Q5.07 referred to the same trip as Q5.06, but rather that the question referred to the longest distance trip that the respondent often made – so for example, the trip home for the holidays, rather than the daily trip to work. We have not adjusted the responses to either question, but suggest that the analyst considers the discrepancy in deciding how to use these responses for a particular purpose.

6 Data recodes & corrections

6.1 Data corrections

Due to the stringency applied to data collection in the field, and extensive advance testing of the data collection tool, minimal data corrections were required. However, for one survey (resp_id 47424), gender and dwelling type were not recorded due to a system error. The appropriate gender (male) and dwelling type (house, brick or concrete structure on a separate stand) were collected from fieldwork records, and used to populate the relevant variables.

All incoming surveys were carefully reviewed for inconsistent responses. Where a single survey contained numerous unexplained inconsistencies, fieldworkers were provided with additional training and supervision, and affected surveys were replaced. In all other instances, we took the decision not to attempt to correct the data, but to accept inconsistencies as an accurate reflection of respondent answers.

Where we have concerns about a particular variable across the dataset as a whole (as documented in Section 5 above), we have not attempted to make any corrections, preferring to allow the analyst to make the most appropriate decision for a particular purpose. We have, however, in some instances provided recodes in addition to the original data. Where recodes are available, these are detailed in Section 6.2 below.

6.2 Data recodes

We provide a number of recodes within the dataset. Many of these are simply to provide more useful analytical categories, while others address concerns with particular variables, as described in Section 5 above. All recodes contain 'recode' in the variable name, and details are provided below:

munic_recode: This simply provides a numerical version of the 'munic' variable, which indicates the municipality in which the interview was conducted.

Table 10: Numeric recode for municipality in which interview was conducted

Original value (munic)	Recode value (munic_recode)
Ekurhuleni	1
Johannesburg	2
Tshwane	3
Emfuleni	4
Lesedi	5
Midvaal	6
Merafong	7
Mogale City	8
Rand West	9

A3_dwelling_recode: This recode simplifies the many categories in A3_dwelling into three main categories - 'Formal', 'Informal' and 'Other'.

Table 11: Details of A3_dwelling_recode

Original value (A3_dwelling)		Recode value (A3_dwelling_recode)	
Value	**Label**	**Value**	**Label**
1	House, brick or concrete structure on a separate stand	1	Formal
2	Traditional dwelling, hut or structure made of traditional materials	3	Other
3	Flat or apartment in a block of flats	1	Formal
4	Cluster house in a complex	1	Formal

5	Townhouse (semi-detached house in a complex)	1	Formal
6	House, flat or room separate from main dwelling in backyard	1	Formal
7	Informal dwelling or shack in backyard	2	Informal
8	Informal dwelling NOT in backyard, e.g. in informal squatter settlement or on a farm	2	Informal
9	Room or flat which is part of main dwelling or property)	1	Formal
10	Caravan or tent	3	Other
11	Unit in a retirement home or barracks etc.	1	Formal
12	Hostel	3	Other
13	Other	3	Other

A3_recode_2: In certain previous iterations of the QoL survey (specifically, QoL I and QoL IV), a 'please specify' field was provided when 'other' was selected for dwelling type. This allowed for a proportion of 'other' responses to be appropriately assigned to one of the standard categories. In QoL V, no further information was collected when 'other' was selected, meaning that none of these responses could be reassigned to pre-existing categories. The result is that QoL V includes a higher proportion of 'other' responses than were present in QoL IV and QoL I. For this reason, a second recode of the dwelling type variable is provided, in which we distinguish between these 'other' responses which might previously have been recoded, and those derived from the selection of 'Traditional dwelling, hut or structure made of traditional materials', 'Caravan or tent', or 'Other'.

Table 12: Details of A3_recode_2

A3_dwelling		**A3_dwelling_recode**	
Value	**Label**	**Value**	**Label**
1	House, brick or concrete structure on a separate stand	1	Formal
2	Traditional dwelling, hut or structure made of traditional materials	3	Other
3	Flat or apartment in a block of flats	1	Formal
4	Cluster house in a complex	1	Formal
5	Townhouse (semi-detached house in a complex)	1	Formal
6	House, flat or room separate from main dwelling in backyard	1	Formal
7	Informal dwelling or shack in backyard	2	Informal

8	Informal dwelling NOT in backyard, e.g. in informal squatter settlement or on a farm	2	Informal
9	Room or flat which is part of main dwelling or property)	1	Formal
10	Caravan or tent	3	Other
11	Unit in a retirement home or barracks etc.	1	Formal
12	Hostel	3	Other
13	Other	4	Other - unspecified

Q1.01.02_people_recode: Given the limited number of households which include 7 or more residents, we provide a recode of household size which collapses all households with 7 or more residents into a single grouping.

Table 13: Details of Q1.01.02_people_recode

Q1.01.02_people	**Q1.01.01_people_recode**	
Value	**Value**	**Label**
1	1	1
2	2	2
3	3	3
4	4	4
5	5	5
6	6	6
7 - 30	7	7+

Q114_recode, Q1.15_recode, and Q116_recode: As described in Section 5 above, challenges were experienced in the implementation of Q1.14_borehole, Q1.15_rainwater and Q1.16_water_truck, which asked whether respondents obtained water from these sources in addition to their main water source. In particular, the 'already mentioned as main water source' was sometimes not selected when it should have been, and in other instances was selected when it should not have been, relative to the responses provided in Q1.07. These recodes are based on changes we were able to make with a fair degree of confidence. Analysts are encouraged to interrogate the original responses and these recodes closely in order to decide whether and how best to use this data. The changes made in generating the recodes we provide are as follows:

- When 'already mentioned as main water source' was selected for all three of these questions, this was replaced with 'no' unless the respondent had specified a particular source as the main water source. In this case, the option specified as the main source was left as 'already mentioned as main water source', while the other two were replaced with 'no'.
- When 'already mentioned as main water source' was selected for one or two of these questions, this was replaced with 'yes' unless the respondent had specified a particular source as the main water source. In this case, the option specified as the main source

was left as ‘already mentioned as main water source’, while the other (if applicable) was replaced with ‘yes’.

- When the respondent selected ‘yes’ or ‘no’ for a water source which was selected as the main water source in Q1.07, this was replaced with ‘already mentioned as main water source’.

Q1.25_recode: Q1.25_1 through to Q1.25_8 each record a yes/no response as to whether the respondent obtains electricity through a particular type of connection. We provide a recode which uses the responses to Q1.25_1 through to Q1.25_9 to provide an indicator as to whether the respondent has access to electricity. This variable is coded 1 (‘yes’) if the respondent responded ‘yes’ to one or more of Q1.25_1 to Q1.25_8. The variable is coded 0 (‘no’) if the respondent responded ‘yes’ to only Q1.25_9, or did not respond ‘yes’ to any of the questions. It should be noted that the recode does not provide an indication of whether the respondent has access to formally provided electricity, as even a connection from a neighbour’s house, from a car battery, or from elsewhere, was considered adequate to code Q1.25_recode as 1.

Q3.01_recode: This recode provides a simplification of the responses in Q3.01_birth_place.

Table 14: Details of Q3.01_recode

Q3.01_birth_place		**Q3.01_recode**	
Value	**Label**	**Value**	**Label**
1	Gauteng	1	Born in Gauteng
2	Eastern Cape	2	Migrated into Gauteng from a province in South Africa
3	Free State	2	Migrated into Gauteng from a province in South Africa
4	KwaZulu Natal	2	Migrated into Gauteng from a province in South Africa
5	Limpopo	2	Migrated into Gauteng from a province in South Africa
6	Mpumalanga	2	Migrated into Gauteng from a province in South Africa
7	Northern Cape	2	Migrated into Gauteng from a province in South Africa
8	North West	2	Migrated into Gauteng from a province in South Africa
9	Western Cape	2	Migrated into Gauteng from a province in South Africa
10	Country outside South Africa	3	Migrated into Gauteng from another country

Q5.01_recode: Respondents had two opportunities to provide the purpose of their most frequent trip. Most respondents provided the purpose in Q5.01_frequent_trip. However, when a respondent indicated that they don’t make any trips, they were asked to confirm this in Q5.02_non_movement. If a respondent indicated in Q5.02_non_movement that they did in fact

make some trips, they were asked to provide the purpose of the most common trip in Q5.02.01_trip. Q5.01_recode replaces “I don’t make any trips” in Q5.01_frequent_trip with the trip purpose as provided in Q5.02.01_trip for these respondents. It should therefore be used as the most complete

Q11.03_recode: This recode provides a categorical version of Q11.03_age, which contains the age of the respondent’s business in years.

Table 15: Details of Q11.03_recode

Q11.03_age	**Q11.03_recode**	
Value	**Value**	**Label**
0 – 1	1	Up to 1 year
2	2	2 years
3 – 4	3	3-4 years
5 – 6	4	5-6 years
7 – 10	5	7-10 years
11 - 15	6	11-15 years
16 - 60	7	16+ years

Q13.01_recode: This recode uses the responses to Q13.01_1 through to Q13.01_10 to provide an indicator as to whether the respondent has participated in any community groups or organisations in the past year. If the respondent responded ‘yes’ to one or more of Q13.01_1 through to Q13.01_1, Q13.01_recode is coded 1 (‘yes’). If the respondent responded ‘no’ to all of these questions, the recode is coded 0 (‘no’).

Q15.01_education_recode: This recode provides a simplification of the responses in Q15.01_education (highest level of education attained).

Table 16: Details of Q15.01_education_recode

Q15.01_education		**Q15.01_education_recode**	
Value	**Label**	**Value**	**Label**
1	No education	1	No education
2	Grade 0 or Grade R	2	Primary only
3	Grade 1 or Sub A	2	Primary only
4	Grade 2 or Sub B	2	Primary only
5	Grade 3, Std 1 or L1	2	Primary only
6	Grade 4, Std 2 or L2	2	Primary only
7	Grade 5, Std 3 or L3	2	Primary only
8	Grade 6, Std 4 or L4	2	Primary only
9	Grade 7, Std 5 or L5	2	Primary only
10	Grade 8, Std 6, L6 or Form I	3	Secondary incomplete
11	Grade 9, Std 7, L7 or Form II	3	Secondary incomplete
12	Grade 10, Std 8, L8, Form III, NTC 1 or RCE higher	3	Secondary incomplete
13	Grade 11, Std 9 or Form IV	3	Secondary incomplete

14	Grade 12, Std 10, Matric	4	Matric
15	A certificate from a college, technikon or university	5	More
16	A diploma from a college, technikon or university	5	More
17	Technikon or university degree	5	More
18	Post graduate degree – e.g. Hons, MA, PhD	5	More
19	Unspecified	6	Unspecified

Q15.02_age_recode: This recode provides a categorical version of Q15.02_age (the respondent's age in years).

Table 17: Details of Q15.02_age_recode

Q15.02_age	**Q15.02_age_recode**	
Value	**Value**	**Label**
18-19	1	18-19
20-24	2	20-24
25-29	3	25-29
30-34	4	30-34
35-39	5	35-39
40-44	6	40-44
45-49	7	45-49
50-54	8	50-54
55-59	9	55-59
60-64	10	60-64
65-105	11	65+

7 GIS coordinates and survey location

7.1 Collection of GIS coordinates

During data collection, up to 29 sets of coordinates were automatically captured for each survey. All data collection devices were set up in advance to rely on GPS satellites to record accurate coordinates, rather than less accurate cell-phone tower triangulation, resulting in high levels of accuracy. Coordinates were regularly reviewed and were invaluable in ensuring appropriate data collection practices.

For each survey, a first ‘site’ coordinate was taken when the survey was opened. The final location was recorded when the fieldworker completed the ‘questions for the fieldworker’ section at the end of the survey. Both of these sets of coordinates were deemed less representative of the survey location than those collected during the survey itself. Site coordinates were often in the street or at the gate of a dwelling, or at the entrance to a complex, depending on the location of the data collector when negotiating access. Final coordinates were often recorded once the data collector had already left the dwelling, as the data collector would not wish to keep the respondent waiting while doing this section. Early in the data collection period, a number of site coordinates were compromised when they were accidentally overwritten by a bug on the data collection system. Consequently, these site and final coordinates were only used in determining survey location in the absence of other location data, and even then with great caution, and only if they were located within the Gauteng province.

Although the data collection instrument was set up to collect coordinates regularly throughout the duration of the survey, this was not always possible. In particular, when data collectors were indoors and away from windows, coordinate accuracy was too poor to record. In addition, for a brief period of time, some data collectors accidentally disabled the geo-coordinate recording functionality during the survey when adjusting device settings to save battery. Consequently, variable numbers of coordinates were available for each survey, and in some instances it was necessary to use the site coordinates, and very occasionally to generate coordinates based on the survey address and photo of the survey location.

7.2 Determination of survey location

All work in aggregating available coordinates to provide a final survey location was done using the WGS1984 geo-projection. In many instances, when the data collector had good satellite access, and did not move during the survey, all coordinates were identical. In these instances, the first set of coordinates other than the site coordinate was recorded as the survey location. In most surveys, however, there was some amount of jitter, and coordinates were close to each other, but not identical. In these cases, when all coordinates were in reasonable proximity and the data was otherwise unproblematic, all available coordinates (excluding the site and final coordinates, and any coordinates outside of Gauteng) were aggregated using the ‘Median Center’ function in ArcGIS. The median coordinates were used as the survey location in these instances.

When no survey coordinates were available, the site coordinates were used, subject to careful checks to ensure that they corresponded to the recorded survey address, photo of the survey location, and the location at which the interview was supposed to have been conducted.

Finally, there were a small number of surveys with no survey coordinates, and a compromised site coordinate. In these cases, if the recorded address and photo of the survey location

corresponded with the location at which the interview was supposed to have been conducted, this target location was used.

Surveys which could not be accurately located were not included in the final dataset.

7.3 Incorporation of spatial paradata

Once a survey location was determined, a spatial merge in ArgGIS was used to locate the survey within the appropriate municipality, ward, main place (MP), sub place (SP), and enumerator area (EA). The exact GIS coordinates for each survey were then removed from the dataset, to protect respondents' anonymity and confidentiality. These coordinates are not made available to researchers.

A number of surveys were found to have been done in a ward or municipality other than that in which they were supposed to have been conducted. This typically happened when a survey location was very near to a ward or municipal boundary, and the data collector inadvertently completed the survey at a dwelling on the other side of the boundary. These surveys were not discarded, but were allocated to the ward in which they were located, and an additional survey was required in the target ward. This was most prevalent in the extremely small inner-city wards and townships.

8 Weighting

8.1 Overview

The random selection of households and individual respondents within each ward was designed to yield a sample matching the adult population of each ward in terms of population group and gender. However, as is always the case, the attained sample did not perfectly reflect the population. In addition, the sample size for each ward was not proportional to ward population, but based on a minimum of 60 interviews for wards in City of Ekurhuleni and City of Johannesburg, and 35 in all other wards. A process of weighting was therefore used to bring the sample to the appropriate distribution by population group and gender within each ward, and to adjust the ward-level sample sizes to the appropriate proportion of the provincial population.

Table 18 provides an overview of the attained sample, unweighted and weighted, in relation to the latest available population data (Census 2011, updated using Community Survey 2016). Table 19 illustrates the distribution of the sample, unweighted and weighted, across Gauteng's nine municipalities, relative to their populations.

Table 18: Sample (unweighted and weighted) and population, by gender and population group

Population group	Gender	Sample %	Census 2011 updated by CS 2016 %	Weighted sample %
African	Male	39.2%	40.0%	40.3%
	Female	45.0%	38.0%	38.4%
Coloured	Male	1.5%	1.6%	1.5%
	Female	2.1%	1.8%	1.8%
Indian/Asian	Male	0.8%	1.4%	1.3%
	Female	0.7%	1.3%	1.3%
White	Male	5.0%	7.3%	7.0%
	Female	5.3%	8.0%	7.7%
Other	Male	0.2%	0.4%	0.4%
	Female	0.1%	0.3%	0.3%

Table 19: Distribution of sample, weighted and unweighted, across municipalities

Municipality	Sample %	Census 2011 updated by CS 2016 %	Weighted sample %
Ekurhuleni	25.3%	25.9%	25.9%
Johannesburg	31.6%	36.6%	36.6%
Tshwane	17.4%	24.1%	24.1%
Emfuleni	6.9%	5.3%	5.3%
Lesedi	1.9%	0.8%	0.8%
Midvaal	2.1%	0.8%	0.8%
Merafong	4.1%	1.4%	1.4%
Mogale City	5.6%	2.9%	2.9%
Rand West	5.2%	2.0%	2.0%

Tables 18 and 19 illustrate quite clearly the ways in which the original distribution of the sample differs from the distribution of Gauteng's population. In particular, African females are somewhat over-represented, while Indian/Asian and white respondents of both genders are under-represented. In addition, as ward level samples were not based on population size, there is some under-representation of larger municipalities. Tshwane in particularly is under-represented due to the smaller minimum ward-level sample size in that municipality (n=35) relative to the other two metropolitan municipalities (n=60). Conversely, the smaller and less densely populated district municipalities are over-represented in the sample. Consequently, population-based weighting is essential in order to ensure that the results of analysis appropriately represent the population of Gauteng as a whole.

8.2 Population estimates used for weighting

Weighting to the ward level requires population estimates broken down by population group and gender for each ward. The latest official statistics at this level are from Census 2011. Given the age of these figures, and the known population growth and demographic shifts since this time, we explored a range of updated population estimates. Data from a range of sources was scrutinised, but no pre-existing estimates were suitable for the purpose of ward-level weighting. Consequently, we used the 2016 Community Survey data, at the municipal level, to update the 2011 ward-level population figures.

Census 2011 figures at the municipal level, for adults, by gender and population group, were used as the starting point (see Table 20). The same table was prepared using the 2016 Community Survey data (see Table 21). For each cell, the ratio between 2011 and 2016 was calculated (see Table 22). Note that as the Community Survey data did not include an 'other' category for population group, the average male and female ratios for each municipality were used for this group.

Table 20: 2011 Census figures (18years +)

	African		Coloured		Indian/Asian		White		Other		TOTAL
	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	
Ekurhuleni	918976	835839	26793	30118	25830	24054	189416	203631	10066	5370	2270093
Johannesburg	1238524	1201287	78739	89594	79679	80171	207784	228111	16476	11846	3232211
Tshwane	781522	778828	19146	21925	20414	19299	220177	243918	9420	5986	2120635
Merafong	68803	52090	702	698	248	159	8802	9284	368	145	141299
Mogale City	100906	92282	1026	1124	2999	2541	28115	30450	906	479	260828
Rand West	78044	65815	4697	5341	479	258	14058	15057	858	340	184947
Emfuleni	204268	215425	2734	2959	2687	2348	33280	35232	1907	753	501593
Lesedi	26778	24608	519	337	577	416	7429	7612	410	83	68769
Midvaal	20656	17367	517	538	258	254	14190	14524	218	164	68686
Gauteng	3438477	3283541	134873	152634	133171	129500	723251	787819	40629	25166	8849061

Table 11: 2016 Community Survey based projection figures (>18years) (http://superweb.statssa.gov.za/webapi/jsf/tableView/tableView.xhtml)

	African		Coloured		Indian/Asian		White		TOTAL
	Male	Female	Male	Female	Male	Female	Male	Female	
Ekurhuleni	1040470	943299	29817	33508	26213	24038	187873	193282	**2478499**
Johannesburg	1396263	1354998	88542	104694	78760	79072	185841	206435	**3494604**
Tshwane	890019	876626	19541	24856	20155	18800	207668	243079	**2300745**
Merafong	62843	49244	665	726	343	213	11213	11396	**136642**
Mogale City	109557	97493	1148	1200	2865	2820	30374	34272	**279728**
Rand West	81308	67996	5374	6108	403	225	14357	15105	**190876**
Emfuleni	211134	216132	3208	3312	2529	2589	32979	37598	**509482**
Lesedi	31480	27890	343	290	388	237	8672	8881	**78180**
Midvaal	22314	18356	746	693	509	431	18974	18879	**80901**
GAUTENG	**3845387**	**3652033**	**149384**	**175387**	**132164**	**128425**	**697950**	**768928**	**9549658**

Table 22: Ratio between Census 2011 and 2016 Community Survey figures

	African		Coloured		Indian/Asian		White		Other	
	Male	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**	**Female**
Ekurhuleni	1.13	1.13	1.11	1.11	1.01	1.00	0.99	0.95	1.05	1.04
Johannesburg	1.13	1.13	1.12	1.17	0.99	0.99	0.89	0.90	1.02	1.03
Tshwane	1.14	1.13	1.02	1.13	0.99	0.97	0.94	1.00	1.02	1.04
Merafong	0.91	0.95	0.95	1.04	1.38	1.34	1.27	1.23	1.18	1.18
Mogale City	1.09	1.06	1.12	1.07	0.96	1.11	1.08	1.13	1.04	1.09
Rand West	1.04	1.03	1.14	1.14	0.84	0.87	1.02	1.00	0.98	0.98
Emfuleni	1.03	1.00	1.17	1.12	0.94	1.10	0.99	1.07	1.02	1.08
Lesedi	1.18	1.13	0.66	0.86	0.67	0.57	1.17	1.17	0.87	0.86
Midvaal	1.08	1.06	1.44	1.29	1.97	1.70	1.34	1.30	1.56	1.41

The ratios in Table 22 were then applied to the Census 2011 data, by population group and gender, at the ward level. This makes the assumption that all wards within a municipality grew at the same rate, which is unlikely in reality, but a more appropriate approach could not be identified. The updated population estimates for Gauteng that were created through this process exceeded the 2016 Community Survey figures by number of individuals in the ‘other’ population group category.

8.3 Implementation of weighting

An iterative re-weighting or ‘raking’ procedure (Battaglia et al, 2004) was used to simultaneously adjust the sample as closely as possible to the appropriate ward, population group and gender categories. The procedure is applied by first weighting the columns (population group and gender) of the table of survey results and then the rows (ward level population size) and then checking whether the discrepancies in the representativeness of the columns is acceptable (since the rows were last weighted, they will correspond exactly to the updated Census 2011 ward figures). After three iterations of the re-weighting procedure the match of the numbers in the race/gender categories to the census figures were considered acceptable, as indicated in the last column of Table 18.

The final weights are presented in Annexure A, and are included in the dataset as the ‘weight’ variable. The weight is automatically applied (‘turned on’) in the SPSS version of the dataset disseminated by the GCRO. However, analysts are advised to ensure that the weights are applied during any analysis in any software, unless there is an explicit intention to use unweighted data. The “Weighted row” column of Annexure A confirms that all wards are now correctly matched to the updated Census 2011 figures.

Reference

Battaglia MP, Izrael D, Hoaglin DC, and Frankel MR (2004), “Tips and Tricks for Raking Survey Data (a.k.a. sample balancing)”, Proceedings of the American Association for Public Opinion Research, Pp 4740 – 4745. www.amstat.org/sections/srms/Proceedings/y2004/.../Jsm2004-000074.pdf