Quality of Life Survey 6 2020/21

QUALITY OF LIFE SURVEY 6 (2020/21)

DATA REPORT

NOVEMBER 2021

Authors:
Sthembiso Pollen Mkhize, Julia de Kadt
and Christian Hamann

GCRO | Gauteng City-Region Observatory

Quality of Life 6 (2020/21) survey: Data report

Authors: Dr Julia de Kadt, Sthembiso Pollen Mkhize, and Christian Hamann

Date: November 2021

Type of output: Technical Report

Research theme: Understanding Quality of Life

Cover Image: Clive Hassall

Published by the Gauteng City-Region Observatory (GCRO), a partnership of the University of Johannesburg, the University of the Witwatersrand, Johannesburg, the Gauteng Provincial Government and organised local government in Gauteng (SALGA).

Suggested citation: de Kadt, J., Mkhize, S.P. and Hamann, C. (2021). *Quality of Life Survey 6 (2020/21): Data report*. Gauteng City-Region Observatory (GCRO). Johannesburg.

TABLE OF CONTENTS

1 INTRODUCTION

This report provides a brief overview of the GCRO's Quality of Life Survey 6 (2020/21) dataset, and describes aspects of the data collection process which have a direct bearing on the structure of the analytical dataset, and the quality of survey data. It also documents variable coding and recoding, derived variables included in the dataset, and implementation challenges which many be relevant to the interpretation of survey data. The report should be reviewed in conjunction with the survey dataset, the questionnaire, and the accompanying technical documentation. This includes the *Field report* (GeoSpace International, 2021), *Sample design* (Hamann & de Kadt, 2021), and *Weighting repo*rt (Neethling, 2021). Please note that unless otherwise specified, all figures provided in this report are based on the unweighted data.

1.1 Overview of dataset

The QoL 6 (2020/21) dataset comprises responses from 13 616 adults, sampled from all 529 wards in Gauteng province. Further detail on sampling processes is available in the *Sample design* report (Hamann & de Kadt, 2021). In each ward, a minimum of 20 adults were interviewed, and in each municipality, a minimum of 600. Details on the composition and distribution of the final sample are available in the Introduction to the *Overview Report* (de Kadt et al. (eds), 2021). Data was collected by GeoSpace International through in person interviews, from late October 2020 through to May 2021. Some high level information on data collection is provided in Section 2 of this report, and more detailed information on data collection is available in the *Field Report* (GeoSpace International, 2021).

For most purposes, the weighted analyses of this dataset are most appropriate. Two weighting variables are included in this dataset. The first, 'DOWNSCALE_MUN_PP_BENCHWGT' is an individual weight, benchmarked to ward level adult population size, and municipal population by race and gender, and then downscaled to the sample size. The second, 'HH_WEIGHT' is a household level weight (not downscaled). Further information on the calculation of these weights is available in the *Weighting repo*rt (Neethling, 2021), and details on appropriate use of the weights is available in the *Guide to weighted data analysis* (Neethling, 2021).

2 DATA COLLECTION BACKGROUND

2.1 Questionnaire development, structure and administration

The survey questionnaire was designed by the GCRO, with input from academic experts, and stakeholders in provincial and municipal government, in late 2019. In response to recommendations from the Quality of Life Survey 10 year review (Orkin, 2020), the questionnaire is shorter than that used in recent iterations of the survey. Nonetheless, the retention of core questionnaire content facilitates longitudinal comparisons and analyses across QoL survey iterations. In addition, new content has been included on social mobility, the impact of COVID-19, and experiences of violence, including intimate partner violence. Content in other areas has been enriched. These are governance, political and governance perceptions, hunger and food security, and environmental risk and vulnerability.

The final questionnaire includes 214 questions, divided into 15 modules, as follows:

1) Dwelling and household information;
2) Access to and satisfaction with basic services;
3) Moving home and migration;
4) Neighbourhood or community;
5) Transport;
6) Household socio-economic status;
7) Governance and government satisfaction;
8) Social and political attitudes;
9) Life satisfaction;
10) Crime and safety;
11) Social and civic participation;
12) Health;
13) Impact of COVID-19;
14) Demographic details; and
15) Exposure to violence, including GBV.

Full questionnaire content is available in the PDF version of the questionnaire which accompanies this report.

2.1.1 Questionnaire administration

Following participant selection and the informed consent process, Sections 1 through 14 were administered by a trained interviewer in a face-to-face interview. Responses were captured on a tablet. Due to the sensitive nature of questions in Section 15, participants were asked whether they were willing to complete Section 15. Those who were willing to do so, self-completed this section on the data collection tablet, although a small proportion did request assistance from the interviewer. Please note that the responses collected in Section 15, with the exception of household income, are not included by default in the public dataset due to their sensitivity. Analysts wishing to use this data must request it separately via the DataFirst website, with a proposal outlining their intended use.

In the PDF questionnaire accompanying this report, the English language question text for each variable is shown in standard font in the 'English Questions' column of the questionnaire PDF, and any additional information shown to the fieldworker is displayed in italics. Response options are provided in the 'English Responses' column.

2.2 Piloting and translation

A process of behind-the-glass piloting of the draft questionnaire, in English, was undertaken in January 2020, which allowed for testing and refinement of content. Following this, the questionnaire was translated into isiZulu, isiXhosa, Setswana, Sesotho and Afrikaans in early 2020, by professional translators at Translation World. These translations were reviewed by a GCRO researcher fluent in each language, and adjusted where necessary, both to ensure that the meaning of the questions were accurately reflected in each translation, and also to ensure that the language used in translations would be easily understood by residents of Gauteng province. At this point, the COVID-19 pandemic meant that work on survey implementation was temporarily put on hold. This was used as an opportunity to develop, in consultation with academic and governmental experts, a short COVID-19 module.

GeoSpace International was appointed as the Quality of Life survey 6 (2020/21) data collection service provider in early 2020, on the basis of a public tender process. On appointment, the GeoSpace field team

reviewed the survey instrument, along with all existing translations. In consultation with the GCRO, a decision was made to translate the questionnaire into three additional languages (Sepedi, Tshivenda and Xitsonga). GeoSpace contracted The Translations Workbench to translate the full questionnaire into these three additional languages, and also to translate the additional COVID-19 module into all nine languages. These additional translations were reviewed and finalised by GeoSpace team members fluent in each language.

Fieldworkers were trained on the English language questionnaire, as well as the translations relevant to their individual language profiles. During training, and a small scale in-field pilot in September and October 2021, feedback from fieldworkers and pilot participants informed further small adjustments to the questionnaire and translations. Please refer to the *Field report* (GeoSpace International, 2021) for further detail on training and in-field piloting.

All translations were included in the electronic data collection application (see Section 2.3 below). Fieldworkers were able to select the respondent's language of choice at the start of each interview, and were also able to move between languages during the course of the interview if needed. The main language in which the interview was completed is available as variable 'interview_lang'. Please note that in addition to the languages noted above, a small number of interviews were also conducted in IsiNdebele (n=88) and Siswati (n=29). In these instances fieldworkers translated questionnaire content independently while administering the interview.

2.3 Data collection system

GeoSpace International used two software systems for the implementation of the QoL 6 (2020/21) survey - HxGN Smart Census and KoBoToolbox. For purposes of field management, GeoSpace customised the HxGN Smart Census system to meet project needs. On the backend, this supported allocation of pre-selected visiting points to field team members, and provided various monitoring dashboards. When installed on dedicated data collection tablets, the HxGN application provided field team members with navigation to the visiting point, and functionally to complete the in-field component of sample selection. Satellite-based GPS coordinates were recorded at various times during the use of this system in field, operating independently of network coverage and with generally good accuracy.

The survey questionnaire was digitised and managed using the KoBoToolbox system. The KoBoCollect application was installed on the data collection tablets, and used to administer the interviews. The KoBoCollect application displayed question text on screen for the interviewer to read out, along with the relevant response options, and in some instances additional notes or information for the fieldworker. All content was shown in the selected interview language. During survey completion, interview duration and GPS coordinates for each section of the questionnaire were recorded.

All aspects of both HxGN Smart Census and KoBoCollect functioned in the field regardless of network coverage. HxGN Smart Census and KoBoToolbox were integrated such that each record's paradata (including GPS coordinates and sampling details) was linked with the relevant interview data. Further information on these systems is available in the *Field report* (GeoSpace International, 2021).

2.3.1 Collection of spatial data

As noted, a number of GPS coordinates were collected for each completed interview. These include the coordinates of the visiting point in the database from which visiting points were sampled; the coordinates captured during the in-field sampling processes, and the coordinates captured during and after interview

completion. Coordinates captured during in-field sampling were generally accurate, as these were always captured outdoors at the interview location. In areas such as complexes, estates and hostels, the sample selection coordinates were sometimes affixed to the entrance location, rather than the selected dwelling unit, making these slightly less reliable. Coordinates captured during interviews were sometimes less accurate or missing altogether when interviews were conducted indoors, due to challenges with satellite connection.

GPS coordinates collected are not shared publicly, in order to protect anonymity and confidentiality. Coordinates were used for Quality Assurance processes, and to generate the spatial paradata, as detailed in section 3.1. Analysts wishing to conduct spatial analyses are encouraged to make use of these spatial paradata variables.

2.4 Interview length

The average duration for the interviewer-administered part of the survey (excluding respondent selection processes) was 34 minutes. Figure 1 presents the distribution of interview length. Based on timed testing of the survey instrument, interviews with durations of less than 15 minutes were deemed unreliable and excluded from the survey dataset. Interviews with durations from 15 to 20 minutes (n=1117) were flagged during Quality Assurance (QA), and manually reviewed prior to a decision on whether they were reliable. Interview duration was also monitored at the fieldworker and team levels, and where these differed notably from average figures this was investigated.

The final dataset includes 5810 interviews (43%) under 30 minutes, of which 8% were between 15 and 20 minutes. No interviews shorter than 15 minutes were approved and included in the final dataset. Only 19 interviews (0.2%) took more than 90 minutes to complete. These were largely caused by survey interruptions and subsequently resumed, but also included some surveys where fieldworkers reported extremely talkative, slow or elderly respondents.

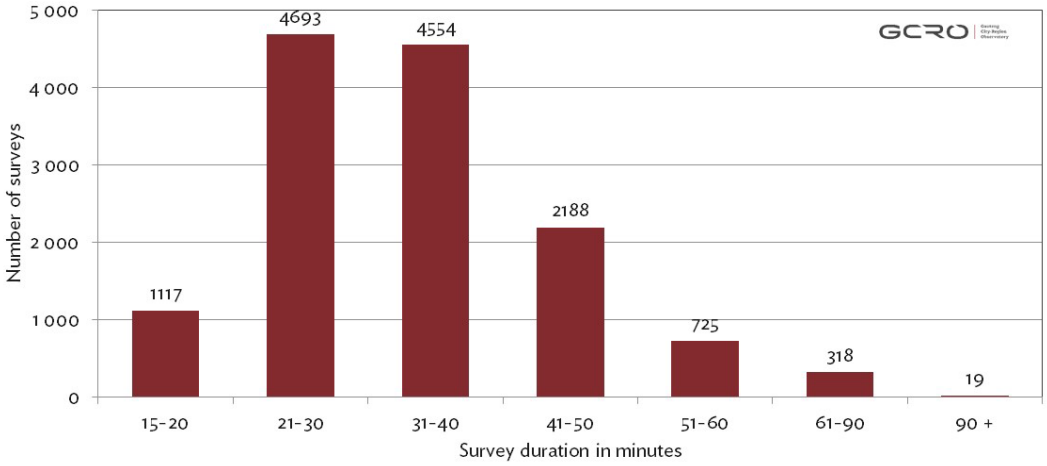

Figure 1: Duration in minutes of surveys in final dataset

Measurement of duration for the self-complete was slightly challenging. In most instances, respondents submitted their responses for the section, as instructed on screen, before returning the tablet to the fieldworker. However, in a few instances, they did not do this, and the fieldworker often only became aware of this when alerted that the module had not been received. In these cases, the duration of this module is

recorded as the time from when the respondent began the module through to the time at which the fieldworker submitted the module. As a result, the self-complete duration was recorded as greater than 20 minutes for almost 3% of those who completed this module. However, reports from the field indicated that it rarely took respondents more than 5 minutes. When duration measures greater than 20 minutes are excluded, the mean duration for this module is 4 minutes, with a standard deviation of 2.3 minutes. Values recorded range from 1 to 20 minutes.

2.5 Question types

2.5.1 Interviewer-administered questionnaire component

All questions in the interviewer-administered component of the questionnaire required the selection of a response, unless they were skipped (see section 2.8 below).The main question types used in the interviewer-administered section of the questionnaire are as follows:

Single response: This is the default question type, offering a set of possible response options, of which only one can be selected. After reading the question to the respondent, the fieldworker waited for the response, and then captured this, repeating the question if necessary. For some questions, the response options were also read to the respondent. Where this was the case, this instruction was displayed on-screen for the fieldworker, and there is a note to this effect in the Notes column of the questionnaire. For some questions, a hard-copy show card was shared with respondents to assist them. Where this was the case, this instruction was displayed on-screen, and is noted in the Notes column of the questionnaire. Many of these questions made use of standardised Likert scales to reduce respondent burden. Details of these scales are included in section 3.3 of this report.

Multi-select: This question type is the same as the single response, with the exception that the respondent could choose more than one response option. These questions are identified as multi-select in the Notes column of the questionnaire. Unless otherwise specified, response options were shared with the respondent, either read out by the interviewer, or through an image-based show card. Examples include Q5.14 (places walked to from home, using a picture show card), Q7.15 (reasons that a bribe had been requested, with options not shared with the respondent), and Q14.3 (disabilities, with options read out to the respondent).

Yes-no list: These questions provide a list of items below the question text, each of which was read to the respondent for a response of yes or no, which was then captured by the interviewer. Examples include Q1.12 (electricity supply), Q1.19 (extreme events) and Q6.3 (household asset list). In Q1.12 (electricity supply), the final two items in the list ('Do not know' and 'No electricity' were only displayed if the respondent had answered 'no' to all previous forms of electricity supply).

A small number of other questions required numerical input (Q3.4 – year moved to Gauteng province; Q5.6 – travel duration; Q5.11 – walking time to public transport; Q14.2 – age; and Q14.5, Q14.6 and Q14.7 – number of household residents). For a number of these variables, the dataset includes a categorical recode, to facilitate analysis. Further information on these variables is available in section 4.1 of this report. Other response types included time (Q5.5 – departure time), free text (Q3.7 – neighbourhood of previous residence; Q5.4.2 – neighbourhood of most frequent trip destination; Q8.25 – what the GCR means to the respondent; Q10.10 – employment occupation; and 'Other – specify' questions as detailed in section 2.7 below), and a dropdown list (Q3.2 - country of birth).

Finally, throughout the questionnaire, there were a number of optional comment boxes allowing free text inputs. These were used by fieldworkers to use if they wished to make a note of any information relevant to

the responses captured, or if the respondent wished to highlight a particular issue for attention. All comments were reviewed during QA and in finalising the analytical dataset, but are not publicly available.

2.5.2 Self-complete questionnaire component

The self-complete section of the interview (Section 15) made use of single response and multi-select question types. In this case, response options were displayed on the tablet for the respondent to read. Multi-select questions included an instruction to select all applicable response options. Where the respondent requested assistance, the fieldworker assisted by reading out the question text and response options. Responses were required to all questions in this section, but a 'Prefer not to answer' option was always available for use where a respondent did not want to provide an answer.

2.6 Coding of free text responses

Free text responses to Q3.7, Q5.4.2, and Q10.10 were reviewed, cleaned and coded as appropriate, during the last few weeks of data collection. Responses to Q8.25 were not cleaned or coded, and are included in their raw form in the dataset for analyst use.

2.6.1 Previous neighbourhood of residence in Gauteng (Q3.7)

Responses to this question were manually checked, and spelling errors were corrected. Where there was uncertainty about a particular response, this was confirmed with the respondent. The name of the neighbourhood was used to link the response to the municipality in which it is located. The municipality name was added to the response to Q3.7. In a small number of instances, there is a discrepancy between the derived municipality included in Q3.7 and the municipality provided in Q3.6a_municipality. No adjustments have been made to address these discrepancies. These errors may be related to changes over time in the municipal location of an area, confusion on the part of the respondent as to the municipality the area fell into, or fieldworker error.

2.6.2 Neighbourhood of destination of most frequent trip (Q5.4.2)

As with Q3.7, responses to this question were manually checked, and spelling errors were corrected. Where responses were unclear, these were confirmed with the respondent. The name of the neighbourhood was used to link it with the municipality in which it is located. The municipality was compared with that selected in Q5.4.1, and where discrepancies were identified, the correct municipal location was confirmed, and the response to Q5.4.1 was adjusted where necessary. Discrepancies were due to confusion on the part of the respondent about the municipality in which the destination fell, fieldworker error, or initial incorrect coding of the response to Q5.4.2.

2.6.3 Employment occupation (Q10.10)

Respondents who had worked in the past 7 days were asked to describe their usual occupation using two or more words. Responses were manually reviewed, and coded in alignment with the major and sub-major occupation codes specified in the South African Standard Classification of Occupations (SASCO) (StatsSA, 2012). Original responses are included in the dataset as variable 'Q10_10_occupation_orig'. Major categories are provided in the variable 'q10_10_occupation_maj', and sub-major in the variable

'q10_10_occupation_submaj'.In a small number of instances (n=11) a respondent was unwilling to disclose their occupation. In both of the coded variables, these responses are coded as refusals.

2.7 Use of 'Other' and 'Other (specify)' response options

Some single response and multi-select questions include an 'Other' response option for use when the standard options are not appropriate. Fieldworkers were trained on how to minimise the use of 'Other' fields by asking for more detail if necessary.

An 'Other (specify)' option was available in a subset of questions was selected on the basis of feedback from our pilot, or high proportions of 'Other' responses in previous survey iterations, for the an 'Other (specify)' option was used. If an 'Other (specify)' response option was selected, a free text field was displayed, and the fieldworker captured further information as provided by the respondent. Towards the end of data collection, free-text responses in 'Other (specify)' fields were reviewed by dedicated QA team members. Where appropriate, these responses were recoded into one of the existing categories. For a number of questions, the details provided suggested the creation of one or more additional response categories. These instances were discussed with the full project team, and a decision was taken as to whether the additional categories should be created. Where this was agreed on, the questionnaire includes in red text "Coded responses based on other (specify)" and new categories are listed here. Where the details did not merit recoding into an existing category, or where numbers were insufficient to introduce a suitable new category, the response remains coded as 'Other (specify)' in the analytical dataset. For the protection of anonymity and confidentiality, we are not able to share the free text responses attached to 'Other (specify)' responses. Details of response patterns and recoding are listed in Section 2.7.2 below.

2.7.1 Use of 'Other'

Table 1 below lists all questions in which an 'Other' response option was available. It also provides the percentage and count of those answering the question who selected 'Other'. In most instances, well below 2% of respondents answering any question made use of 'Other'. There were two questions in which more than 2% responded with 'Other', however. In Q13.2 on why the respondent doesn't make use of public health care services, 10% of the respondents who were asked this questions used 'Other', and in Q14.9 on the language most frequently used in the home, 3% of respondents selected 'Other'.

Question	Percentage (count) 'Other'
Population group (A1)	0.1% (n=17)
Dwelling tenure (Q1.3)	1.1% (n=150)
Ownership type (Q1.3a)	*1.2% (n=100 of 8 456)
Rental type (Q1.3b)	*2.0% (n=64 of 3 238)
Main water source (Q1.4)	0.3% (n=42)
Main toilet type (Q1.10)	0.2% (n=25)
Refuse removal (Q1.11)	1.1% (n=145)
Energy for cooking (Q1.15)	0.1% (n=16)
Energy for lighting (Q1.16)	0.2% (n=31)
Type of dwelling moved from (Q3.8)	*0.9% (n=92 of 9 724)
Municipality travelled to (Q5.4.1)	*0.9% (n=12 of 13 304)
Mode of transport used (Q5.7)	*0.6% (n=79 of 13 304)

Walk to 'other' place (Q5.14)	*1.7% (n=224 of 13 304)
Why don't use public health (Q13.2)	*10.0% (n=368 of 3 684)
Relationship status (Q14.4)	0.5% (n=65)
Main language (Q14.9)	3.0% (n=408)

Table 1: Availability and use of 'Other' response options across the questionnaire. *Note, percentage figures are the percentage of those answering the question, and not the percentage of the full sample.

2.7.2 Use and recoding of 'Other (specify)'

Table 2, below, lists all questions which made use of an 'Other (specify)' response option. The table provides information on the percentage and count of respondents who originally selected 'Other (specify)', and the percentage and count with answers still coded as 'Other (specify)' in the final analytical dataset following coding of responses. The final column in the table lists new response categories added to the variable, if any.

In some instances the added categories are very small (see Q1.18, Q4.2 and Q7.10). In these instances analysts should decide whether they wish to retain these as separate categories, or treat them as part of the 'Other' category.

Implementation of Q5.1 and Q5.3 was somewhat challenging, and further information is available in sections 4.2.6 and 5.4 below. There are two issues of relevance to Table 2 however. Firstly, while an 'Other (specify)' option was available in both Q5.1 and Q5.3, it was not used by any respondents answering Q5.3. Consequently, Table 2 provides the data only for Q5.1. Responses to both Q5.1 and Q5.3 are consolidated in the variable 'q5_1_3_frequent_trip_cons', and all recodes conducted for Q5.1 are also reflected in this variable. Secondly, two of the new categories in the final column were briefly added to the digital questionnaire, and selected by a small number of respondents. These were 'Family & Friends' (n=65) and 'Health care' (n=7).

Question	**Percentage (count) of 'Other (specify)'**		**New categories generated (percentage and count)**
	Original data	Final data	
Dwelling type (A3)	0.2% (n=33)	0.0% (n=3)	N/A
Electricity generation purpose (Q1.13a)	*2.3% (n=20 of 885)	*1.1% (n=10 of 885)	N/A
Reason for municipal arrears (Q1.18)	*5.5% (n=132 of 2 392)	*2.1% (n=50 of 2 392)	*Dispute over who pays (0.5%, n=11)
Reason moved to neighbourhood (Q3.9)	*6.4% (n=626 of 9 724)	*1.2% (n=113 of 9 724)	N/A
Biggest community problem (Q4.2)	3.8% (n=513)	1.6% (n=221)	Lack of security services (0.0%, n=6)
Travel purpose (Q5.1)	7.1% (n=967)	0.5% (n=63)	Family & friends (3.1%, n=418) Health care (2.0%, n=277) Church (0.9%, n=119)
Reason didn't vote (Q7.3)	*23.2 (n=311 of 1 341)	*7.2% (n=97 of 1 341)	Working (3.7%, n=49) Self/family member sick (6.0%, n=81) Out of province/country (2.3%, n=31)
Government worst at (Q7.10)	4.2% (n=571)	1.8% (n=251)	All of the above (0.4%, n=51)

Reason for bribe (Q7.15)	*12.3% (n=136 of 1 102)	*5.9% (n=65 of 1 102)	Work (4%, n=44)
Reason for poverty (Q8.13)	3.5% (n=481)	2.4% (n=324)	N/A
Other clubs or organisations (Q12.1)	1.0% (n = 131)	**1.0% (n=131)	N/A

Table 2: Availability and use of 'Other' response options across the questionnaire. *Note, percentage figures are the percentage of those answering the question, and not the percentage of the full sample. ** Note, during coding, 'other specify' responses were coded to the correct variables, but the variable for the 'other specify' responses was not recoded to 0. As a result the number of these responses does not change.

2.8 Skip patterns

A number of skip patterns were used in the questionnaire, to avoid asking respondents questions that were not applicable to them. All skip patterns are documented in the survey questionnaire, and are also listed here. Where a particular response option triggers a skip, this is indicated in brackets after the response option. Where a particular question is only asked when particular conditions are met (or inversely, skipped under certain conditions), this is specified in the 'Notes' column. When a question was not asked of a respondent, the dataset makes use of a '-1' to indicate that the response is missing due to a skip pattern. Further information on coding of missing data is available in section 3.2 of this report.

2.8.1 Main questionnaire

Skip patterns were used in the following questions in the main questionnaire:

Address (Q1.1, Q1.2a & Q1.12b): Q1.1 asks respondents whether they have an address. If respondents answered that they did not, or that they did not know, they were not asked for address details (Q1.2a), and were instead asked to describe how to locate their dwelling (Q1.2b). Conversely, those who had an address were not asked Q1.2b. Please note that Q1.2a and Q1.2b are not available for dissemination as they contain identifying information.

Number of households sharing a room (Q1.2c & Q1.2d): Q1.2c asked respondents how many rooms their household occupies. Respondents who indicated 1 room where then asked how many other households share this room, as a measure of crowding. When the household occupied more than one room, Q1.2d was not asked.

Household tenure (Q1.3, Q1.3a & Q1.3b): Respondents were asked about dwelling tenure in Q1.3. If respondents indicated that the dwelling was owned by the household, they were asked for further ownership details in Q1.3a. If respondents indicated that the dwelling was rented, they were asked for further details on the rental arrangement in Q1.3b. When any other option was selected in Q1.3, these additional questions were both skipped.

Main water source (Q1.4, Q1.5, Q1.6 & Q1.7): All respondents were asked Q1.4, about the household's main water source. When respondents indicated that they received piped water, either into the dwelling or into the yard, they were additionally asked about the type of water meter the household had (Q1.5). Those who selected any other response option were not asked Q1.4. Additionally, respondents who answered 'Well or borehole' to Q1.4 were not asked Q1.6 (whether the household also gets water from a well or

borehole), and those who selected 'Rainwater tank (eg. Jojo tank)' were not asked Q1.7 (whether the household also gets water from a rainwater tank).

Electricity supply (Q1.12): This question asked respondents to identify all forms of electricity supply that their household made use of, and was implemented as a yes-no list (see section 2.5.1 above for more information.) Item 2 ('Electricity with post-paid meter) was only displayed and read out if the respondent selected 'No' to item 1 ('Electricity with pre-paid meter'). Additionally, the final two options ('Do not know' and 'No electricity') were only displayed and read out to the respondent if they had answered 'no' to all of the seven preceding items. Respondents who answered 'Yes' to 'No electricity' were next asked Q1.15.

Electricity generation plans for those already generating electricity (Q1.12a & Q1.12b): Q1.12 asked respondents to identify all forms of electricity supply that their household made use of. If a respondent answered 'Yes' to 'Solar, wind or PV power' and/or 'Petrol or diesel generator' in Q1.12, they were identified as generating some of their own electricity, and were asked Q1.12a – whether they were generating electricity for daily use, or only for use during loadshedding or power cuts. If the respondent answered 'loadshedding or power cuts only', they were additionally asked Q1.12b – whether they were planning increase their electricity generation in the coming year. Those who selected 'For daily use' in Q1.12a skipped the remaining electricity generation questions, and were next asked either Q1.14 or Q1.15 as appropriate based on their other responses in Q1.12.

Electricity generation plans for those not generating electricity (Q1.13 & Q1.13a): Respondents who answered 'No' to both 'Solar, wind or PV power' and 'Petrol or diesel generator' were classed as not currently generating any of their own electricity. Unless they had also answered 'Yes' to 'No electricity', these respondents were asked Q1.13 – whether the household was planning to start generating electricity in the next 12 months. Those who answered 'Yes' were additionally asked Q1.13a – whether they were planning to start generating electricity for daily use, for loadshedding and power cuts only, or some other arrangement. Those who didn't answer 'Yes' to Q1.13 were next asked Q1.14 or Q1.15 as appropriate based on their other responses in Q1.12.

Electricity supply interruptions (Q1.14): Respondents were only asked Q1.14, about how frequently they experienced electricity interruptions in the past 12 months, if they had answered 'Yes' to either 'Electricity with pre-paid meter' or 'Electricity with post-paid meter' in Q1.12.

Reason for arrears in water or electricity accounts (Q1.17 and Q1.18): Q1.17 asked respondents whether they had unpaid water or electricity accounts. Respondents who answered 'Yes' to Q1.17 were asked Q1.18 (main reason for arrears), while all others skipped to Q1.19.

Damage caused by extreme events (Q.19 and Q1.20): Respondents who had responded 'Yes' to experiencing one or more of the extreme events listed in Q1.19 were additionally asked Q1.20 (whether the event(s) had caused damage or injury). Q1.20 was not asked of those who responded 'No' to all extreme events.

Place of birth (Q3.1 & Q3.2): In Q.3.1, respondents were asked to identify their province or country of birth. When respondents indicated that they were born outside of South Africa, the name of the country was captured in Q3.2. This question was skipped for those born in South Africa.

Timing of move to Gauteng (Q3.1, Q3.3 & Q3.4): Respondents who indicated in Q3.1 that they were not born in Gauteng were asked about where they last lived before coming to Gauteng (Q3.3), and when they had moved into Gauteng (Q3.4). Those who indicated in Q3.1 that they were born in Gauteng were not asked these questions.

Moving into current neighbourhood (Q3.5): All respondents, regardless of place of birth, were asked about how long they had lived in their current neighbourhood. Those who indicated that they had always

lived in that neighbourhood skipped through to Q3.10, while all other respondents continued to answer additional questions about the move to their current neighbourhood. If respondents who indicated that they were born outside of Gauteng selected 'I've always lived here', a logic check was triggered (see section 2.9.1).

How respondent came to current neighbourhood (Q3.6, Q3.7, Q3.8 & Q3.9): These questions were only asked of individuals who selected a response other than 'I've always lived here' in Q3.5. In Q3.6, respondents were asked whether they'd moved to their current neighbourhood from somewhere else in Gauteng, another province, or another country. Only those respondents who moved from somewhere else in Gauteng were asked Q3.6a and Q3.7, which asked for the details of the municipality and neighbourhood that they moved from. Those who moved to their current neighbourhood directly from another country or province skipped through to Q3.8 (what dwelling type they moved from).

Purpose of most frequent trip (Q5.1-Q5.3): Q5.1 asks respondents the purpose of their most frequent trip. If a respondent answered that they don't make any trips (n=289, or 2.1% of respondents), this was recorded, but they were then asked Q5.2, which asked them if they were sure that they never go anywhere. A small number of respondents (n=8) answered 'no' to this question. These individuals were asked Q5.3, which offered the same response options at Q5.1. We have provided a variable which includes information on all trip purposes, whether captured in Q5.1 or Q5.3 – more information is available in section 4.2.6. Individuals who confirmed in Q5.2 that they really never go anywhere skipped through to Q5.13.

Details of most frequent trip (Q5.4.1, Q5.4.2, Q5.5, Q5.6, Q5.7, Q5.8 & Q5.9): Q5.4.1 through to Q5.9 were only answered by respondents who reported a trip in Q5.1 or Q5.3. Q5.7 asked the respondent for the mode of transport used to cover the longest distance the last time they made their most frequent trip. Those who selected any form of public transport (minibus taxi, train, Gautrain, BRT or any other form of bus) were asked Q5.8 and Q5.9 about public transport safety, while others skipped through to Q5.10.

General transportation information (Q5.10, Q5.11, Q5.12, Q5.13 & Q5.14): Of these questions, only Q5.13 (involvement in a traffic accident) was asked of respondents who reported that they never go anywhere in Q5.2.

Debt (Q6.2 & Q6.2a): All respondents were asked in Q6.2 about whether they were currently in debt. Those who responded 'Yes' were asked Q6.2a, which asked whether they had missed a debt repayment in the past three months. Those who answered 'No' to Q6.2 skipped through to Q 6.3.

Child hunger (Q6.5 & Q6.5a): All respondents were asked Q6.5, which asks whether there has been a time in the past 12 months where there was not enough money to feed children in the household. Response options were 'Yes', 'No' and 'There are no children in the household'. Those who said that there were no children in the household were not asked Q6.5a (whether children benefitted from school feeding schemes), but skipped through to Q6.6. Those who answered 'Yes' or 'No' to Q6.5 were asked Q6.5a before continuing to Q6.6.

Voter registration and electoral participation (Q7.1, Q7.2 & Q7.3): Q7.1 asked respondents whether they were a registered voter. Those who said 'No' skipped through to Q7.4. Those who said 'Yes' in Q7.1 were asked Q7.2 (whether they had voted in the 2019 National elections). Those who said that they had voted then skipped through to Q7.4, while those who said that they had not voted were asked Q7.3 (the main reason that they did not vote).

Being asked for a bribe (Q7.13, Q7.14 & Q7.15): Q7.13 asked all respondents whether they had ever been asked to pay a bribe to a government official, member of the police, or other public servant. Those who said 'No' skipped through to Q8.1. Those who said 'Yes' in Q7.13 were asked Q7.14 (whether they had been asked

for a bribe in the past 12 months). Those who said 'No' in Q7.14 skipped through to Q8.1, while those who said 'Yes' in Q7.14 were asked in Q7.15 to indicate what bribe had been requested for.

Familiarity with the term "Gauteng City-Region" (Q8.24 & Q8.25): These questions were originally only meant to be asked of a small proportion of respondents, but due to implementation challenges (see section 5.8 for more detail), they were asked more widely. Q8.24 asked the respondent to indicate their agreement with the statement "I am familiar with the term "Gauteng City-Region"'. Only those who agreed with the statement ('Strongly agree' or 'Agree') were asked Q8.25, which asked them to explain in their own words what 'Gauteng City-Region' meant to them. Those who did not agree in Q8.24 skipped through to Q9.1.

Currently working (Q10.2, Q10.3): All respondents were asked whether they did any type of work in the past 7 days (Q10.2). Those who said 'Yes' were categorised as working, and skipped through to Q10.6. Those who said 'No' were then asked Q10.3 (whether they been appointed to a new job but had not yet started). Those who said 'Yes' to Q10.3 were additionally categorised as working, and skipped through to Q10.6, while those who said 'No' to Q10.3 were categorised as not working, and were asked Q10.4

Unemployment (Q10.4 & Q10.5): All non-working individuals (see above) were asked if they were unemployed and looking for work (Q10.4). Those who responded 'Yes' were not asked any further questions about work, and skipped through to Q11.1. Those who answered 'No' in Q10.4 were then asked why they weren't looking for work (Q10.5). After answering this question, they then skipped through to Q11.1.

Employment details (Q10.2, Q10.3, Q10.6, Q10.7, Q10.8, Q10.9 & Q10.10): Only the individuals categorised as working (see above) in Q10.2 or Q10.3 were asked questions Q10.6 through to Q10.10, which collected further information about the nature of the respondents work.

Protest in neighbourhood (Q12.3 & Q12.4): All respondents were asked whether there had been a protest of any kind in their neighbourhood or community in the past year (Q12.3). Those who answered 'Yes' were asked in Q12.4 what the protest was about, while all others skipped directly to Q12.5.

Healthcare services used (Q13.1, Q13.2, Q13.3 & Q13.4): All respondents were asked in Q13.1 about where they usually went for healthcare. A small number of respondents answered 'Not applicable, don't usually need health care'. These individuals were not asked any further questions about health services used, and skipped through to Q13.5. Individuals who indicated in Q13.1 that they used only public health care facilities, or a combination of public and private facilities were not asked Q13.2, but skipped through to answer Q13.3 and Q13.4. Those who answered in Q13.1 that they used only private health care facilities, traditional healers or spiritual healers were asked in Q13.2 for the main reason that they don't make use of public health care facilities. They were then also asked Q13.3 and Q13.4.

Social relief of distress grant (Q13.14 & Q13.15): All respondents were asked in Q13.14 whether they or another household member had applied for the R350 COVID social relief of distress grant. Those who responded 'Yes' were then asked in Q13.15 whether any applicant had received the grant. Those who answered 'No' in Q13.14 skipped through to Q13.16.

Number of people in household (Q14.5, Q14.6 & Q14.7): In Q14.5, respondents were asked for the number of people, including babies and children, living in the household. If respondents answered '1', they skipped Q14.6 which asks for the number of residents under the age of 18, as respondents were all aged 18 or more. If the respondent indicated more than one resident, they were asked for the number of residents under 18 in Q14.6, and the number of residents aged 60 or more in Q14.7. Validation was applied to responses, as detailed in section 2.7 below.

2.8.2 Self-complete module

Due to the sensitive nature of the questions in this module, a second consent process was conducted to ensure respondents agreed to participation prior to being offered the tablet to complete this section. If a respondent did not consent to participate in this section, the interview was terminated at this point, and all questions in the self-complete section are set as missing due to skip (-1). The variable 'sc_consent' indicates whether or not the respondent consented to participate in this section.

For respondents who did complete the section, a 'Prefer not to answer' response option was available for every question. This response option is coded differently across variables, as indicated in the questionnaire. In most instances, selection of this option would simply move the respondent to the next question, but in some instances selection of 'Prefer not to answer' triggered a skip across a few related questions, as detailed below.

Skip patterns were used in the following way in the self-complete module:

Being hit (Q15.8, Q15.9 & Q15.10): In Q15.8, participants were asked whether any person had hit them with a fist or something else, or beaten them in the past 12 months. Those who responded 'No' or 'Prefer not to answer' skipped through to Q5.11. Those who responded 'Yes' in Q15.8 were then asked in Q15.9 to select all applicable perpetrators. Those who selected 'Current or former partner' were then also asked Q15.10, unless they had additionally selected 'Prefer not to answer', in which case they skipped to Q15.11. All respondents who did not select 'Current or former partner' also skipped directly to Q15.11.

Being kicked (Q15.11, Q15.12 & Q15.13): In Q15.11, participants were asked whether any person had kicked, dragged, pushed, shoved, choked or burnt them in the past 12 months. Those who responded 'No' or 'Prefer not to answer' skipped through to Q5.14. Those who responded 'Yes' in Q15.11 were then asked in Q15.12 to select all applicable perpetrators. Those who selected 'Current or former partner' were then also asked Q15.13, unless they had additionally selected 'Prefer not to answer', in which case they skipped to Q15.14. All respondents who did not select 'Current or former partner' also skipped directly to Q15.14.

Threatened/harmed by gun or knife (Q15.14 & Q15.15): In Q15.14, participants were asked whether anyone had used or threatened to use a knife or gun against them in the past 12 months. Those who responded 'No' or 'Prefer not to answer' skipped through to Q15.16. Those who selected 'Yes' were then asked in Q15.15 to list all applicable perpetrators, prior to continuing to Q15.16.

Rape (Q15.16, Q15.17 & Q15.18): In Q15.16, participants were asked whether they had sex in the past 12 months because they were physically forced or were afraid they might be physically harmed if they refused. Those who responded 'No' or 'Prefer not to answer' skipped through to Q5.19. Those who responded 'Yes' in Q15.16 were then asked in Q15.17 to select all applicable perpetrators. Those who selected 'Current or former partner' were then also asked Q15.18 unless they had additionally selected 'Prefer not to answer', in which case they skipped to Q15.19. All respondents who did not select 'Current or former partner' also skipped directly to Q15.19.

Please note that the variables in the self-complete module are not included in the publicly available survey dataset, and must be requested separately from DataFirst.

2.9 Logic checks, validations and value constraints

A small number of logic checks and validations were built into the questionnaire, and used during data collection to provide fieldworkers with an opportunity to confirm these responses. A number of fields requiring a numerical response also included built in value constraints, to reduce errors in entering

numbers. These logic checks, validations and value constraints are documented below. As previously indicated, all questions in the main survey required a response, unless skipped by a skip pattern. It was not possible for the fieldworker to continue to the next question unless all previously asked questions had responses recorded.

As part of the QA process, a further series of automated data checks were run on all incoming survey responses to identify anomalous responses. Where these data checks identified issues, surveys received additional manual scrutiny from the QA team. These automated checks are detailed in Section 2.10.

2.9.1 Logic checks and validations built into the questionnaire

The following checks were built-in to the survey questionnaire, and used during data collection:

Electricity supply (Q1.12): In this yes-no list, it was not possible for respondents to respond 'Yes' to both options 1 ('Electricity with pre-paid meter') and 2 ('Electricity with post-paid meter'). This was achieved by not displaying option 2 unless 'No' was recorded for option 1. Where option 2 was not displayed, this variable is coded as missing due to skip ('-1').

Similarly, it was not possible for respondents to respond 'Yes' to option 8 (Do not know) or 9 ('No electricity') along with any of options 1-7. This was achieved by not displaying items 8 or 9 unless a 'No' had been recorded for all previous items. If items 8 and 9 were not displayed, they are coded as 'missing due to skip' ('-1').

Place of birth and duration of stay in current neighbourhood (Q3.1 & Q3.5): If a respondent indicated in Q3.1 that they were born outside of Gauteng province, and then in Q3.5 indicated that they had always lived in their current neighbourhood, an alert was displayed on the screen to ask the field worker to review these responses. It was not possible to continue until this was resolved.

Most frequent trip (Q5.1, Q5.2 & Q5.3): In Q5.1, if a respondent selected the response option 'I don't make any trips', this was verified by asking the respondent to confirm that they never leave to go anywhere in Q5.2. If the respondent confirmed that this was correct, they skipped all related questions. If a respondent indicated that they did in fact go somewhere, these details were collected in Q5.3, and the respondent completed all related questions.

Places walked to from home (Q5.14): It was not possible to select option 16 ('Nowhere') in combination with any other response options. If this was attempted, an alert appeared on the screen, and it was not possible to progress until this was corrected.

Protein sources (Q6.6): Although this question was asked as a multi-select question, it was not possible to select more than two responses. If more than two were selected, an alert appeared on the screen, and it was not possible to progress until this was correct. It was, however, possible for only one response option to be selected.

Number of household residents (Q14.5, Q14.6 & Q14.7): After answering each of Q14.6 (number of residents under 18) and Q14.7 (number of residents aged 60 or above), respondents were asked to confirm their response prior to continuing to the next question. If the response to either Q14.6 or Q14.7 exceeded the value provided in Q14.5 (total number of household residents), an alert was displayed on the screen, and it was not possible to proceed until this was corrected. Additionally, if the sum of Q14.6 and Q14.7 exceeded the answer provided in Q14.5 an alert was displayed on screen, and it was not possible to continue until this was corrected.

2.9.2 Value constraints

Value constraints were applied to a number of questions requiring numerical responses, to limit the chance of an invalid number being erroneously captured. These constraints are detailed in the questionnaire, and below.

Year moved to Gauteng (Q3.4): The value entered could not be less than 1915. For interviews conducted in 2020, it could not be more than 2020, and for those conducted in 2021, it could not be more than 2021.

Duration of most frequent trip in minutes (Q5.6): The value entered could not be less than 0.

Respondent age (Q14.2): The value entered could not be less than 18, and could not be more than 115. All ages over 80 were verified during Quality Assurance. Please note that only a recode of the age variable is included in the public dataset, in order to protect the anonymity of participants.

Number of people living in household (Q14.5): The value entered could not be more than 30.

Number of people under 18 living in the household (Q14.6): The value could not be more than the response recorded in Q14.5.

Number of people aged 60 or more living in the household (Q14.7): The value entered could not be less than 0, and could not be more than the responses recorded in Q14.5.

2.10 Quality assurance processes

Quality Assurance (QA) processes were conducted by both GeoSpace International and the GCRO. Key sampling, spatial and data-set QA processes are documented here, and further information on GeoSpace International QA is also available in the *Field report* (GeoSpace International, 2021).

The general QA workflow involved a series of automated checks run by GeoSpace International on all incoming interviews, as detailed below. Interviews which were flagged by these automated checks received manual review, which might trigger confirmation of responses with the fieldworker or respondent, a return to field, correction of data errors, or the rejection of the interview altogether. On a weekly basis, GeoSpace International also examined QA results at the level of the individual fieldworker, and occasionally at the team level. The GCRO received completed interviews from GeoSpace International, with a record of QA results and progress for each questionnaire.

The GCRO independently ran a series of similar automated checks, to ensure that the GeoSpace International processes were working effectively. In addition, the performance of skip patterns and logic checks was reviewed on a weekly basis. Periodic manual checks were also used to look for any anomalies in the data, and a proportion of fieldworker and respondent comments were manually scanned. On detection of anomalies, the GCRO notified GeoSpace International, who took responsibility for resolution of the issue.

2.10.1 Sampling and implementation checks

Linkage of HxGN Smart Census and Kobo Toolbox data: GeoSpace International ensured that each record on HxGN Smart Census was linked to exactly one completed questionnaire on Kobo Toolbox, and vice versa. Where this was not the case, data was examined to enable manual matching.

Duplication of interview: GeoSpace International ensured that no records were duplicated, through examination of spatial, sampling and other para-data.

Linkage of main questionnaire and self-complete content: GeoSpace International ensured that there was exactly one valid self-complete record for each valid main questionnaire, and that these were linked correctly. (Note that even when the respondent chose not to complete the self-complete module, the fieldworker was obliged to submit a record of this decision).

Use of substitution points: Where interviews were conducted at substitution points, GeoSpace International ensured that the original pre-selected visiting points had been visited, with appropriate revisits as applicable, and that it had not been possible to conduct an interview at this point. GeoSpace International additionally ensured that the visits to the original visiting point preceded engagements with substitution points. However, towards the end of data collection, when there were few remaining visiting points in a given area, this second requirement was relaxed, so that fieldworkers could assess viability of substitute points while waiting to revisit the original visiting point.

Number of visits per EA: GeoSpace International continuously monitored the number of interviews per EA, to ensure that the sample within each ward was appropriately distributed. This was particularly important in areas where there was extensive use of substitution.

Interview location: Throughout the data collection period, GeoSpace International validated the location of all interviews by ensuring interview coordinates were aligned with sample points. Where this was not the case, the interview was investigated. In many instances discrepancies could be explained by the distance between the point at which an interview could be conducted, and the original sample point. Where the difference could not be explained, the interview was not accepted. Interview location was also validated by calculation of the distance between coordinates collected throughout the interview. If there were notable distances between coordinates, this required investigation. GCRO independently replicated interview location validation at various points during the course of data collection.

Appropriate respondent selection: The name, age and sex of the sampled respondent in the household register was compared to information collected during the interview itself. Discrepancies were investigated. Interviews conducted with individuals other than the sampled adult were not accepted.

Interview language: Interview language was compared to reported home language. Where this differed, the fieldworker was asked to provide an explanation in a comments field. Explanations were manually reviewed.

Interview duration: Interviews with a main questionnaire duration of less than 20 minutes were all examined manually. Interviews with a main questionnaire duration of less than 15 minutes were rejected.

2.10.2 Questionnaire content checks

Number of households sharing a room: Where the respondent reported in Q1.2c that the household lives in a single room, and further indicated that the household shares this room with other households in Q1.2d, this information was verified, particularly if the number of other households was identical to the number of household members.

Number of adult residents: The number of adults listed in the household listing was validated against the number of adults indicated by responses to Q14.5 (total number of residents) and Q14.6 (total number of children).

Fieldworker comments: All interviews which included fieldworker comments were flagged, and the comments were manually reviewed by the GeoSpace International QA team. If necessary, further information was obtained from the respondent or fieldworker.

2.10.3 Automated fieldworker and team level checks

A number of checks were also examined at the fieldworker and team levels, as follows.

Duration: Number of interviews shorter than 20 minutes, average questionnaire duration, and distribution of durations were examined at the fieldworker level. Where fieldworkers had a high proportion of very short questionnaires, or unusual average duration or distribution of durations, this was investigated further, and remedied as appropriate.

Household listing: Average number of adults listed for each household was examined, to ensure that fieldworkers were listing adult residents appropriately. Where the average number differed notably from the mean across all fieldworkers, this was investigated further, with the provision of further training if needed.

Respondent sex: The balance of male and female interviews conducted by each interviewer was monitored, and where this varied notably from an even split, this was investigated. In some instances, there were legitimate explanations for this, such as area demographics or fieldworker skill sets.

Refusal rates for self-complete module: The proportion of respondents refusing to complete the self-complete module was monitored. Where this differed substantially from the mean, this was examined. Further training was provided to fieldworkers who had very high refusal rates.

Refusal rate for provision of contact information: The proportion of respondents refusing to provide contact information was monitored. Where this differed substantially from the mean, this was examined, and training offered if appropriate.

3 PARADATA, STANDARDISED CODES AND RESPONSE OPTIONS

This section provides an overview of the paradata included with the QoL 6 data, and documents the standardised codes used for missing data and frequently used response options in the dataset.

3.1 Spatial and other paradata

Paradata refers to variables in a survey dataset which describe the data collection process. For reasons of confidentiality, not all paradata related to the QoL 6 survey can be made publicly available. Where possible, paradata variables are included in their raw form, but in other instances we are only able to provide derived paradata variables.

3.1.1 Determination of survey location

Survey location was determined on the basis of the GPS coordinates captured when the adult roster was completed, prior to the selection of the respondent. These coordinates were used as they corresponded to

the respondent's actual place of residence, and generally had good accuracy. These coordinates were used in the generation and validation of all spatial variables included in the dataset.

3.1.2 Paradata variables included with the dataset

The paradata variables included in the QoL 2020/21 dataset are listed and described in the table below.

Variable name	Description
unique_id	Unique identifier for each respondent, automatically generated by data collection system during data collection.
interview_date	Full calendar date on which the interview was conducted.
municipality	Text variable providing the name of the Metropolitan or Local municipality in which the interview was conducted, as recorded during data collection and subsequently validated using interview geo-coordinates.
District_municipality	Numerically coded variable providing the Metropolitan or District municipality in which the interview was conducted, generated using the 'municipality' variable.
municipality_coded	Numerically coded variable providing the Metropolitan or Local municipality in which the interview was conducted, generated using the 'municipality' variable.
Planning_region	Text variable. For interviews in Metropolitan municipalities, this variable provides the municipal planning region, generated using interview geo-coordinates. For interviews in Local municipalities, the Local municipality name is provided.
Planning_region_code	Text variable providing planning region or local municipality codes.
ward_code	Ward in which the interview was conducted, recorded during data collection and subsequently validated using interview geo-coordinates.
ea_code	Enumeration Area in which the interview was conducted, generated using interview geo-coordinates.
adult_count	Number of adult residents listed during the household listing prior to random selection of the respondent.
dur_mins	Duration of the interviewer-administered questionnaire component in minutes, as recorded by the data collection system.
interview_lang	Interview language selected by the respondent, as recorded by the fieldworker at the beginning of the interview.
sc_consent	Whether the respondent consented to the self-complete questionnaire component.

Table 3: Paradata included with the QoL 6 dataset

3.2 Missing information

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

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

Table 4: Standard codes of missing data

Information on which questions make use of skip patterns, and include missing data due to these skip patterns is available in section 2.8. More information about the missing information in the section 5 questions is available in section 5.4 of this report.

The following variables have data missing due to a fieldwork error:

- Q3.2: Country of birth (n=40)
- Q5.4.1: Municipality of destination for most frequent trip (n=33)
- Q5.4.2: Neighbourhood of destination for most frequent trip (n=33)
- Q5.5: Time of departure for most frequent trip (n=33)
- Q5.6: Duration of most frequent trip (n=33)
- Q5.7: Mode of transport for most frequent trip (n=33)
- Q5.8: Safety waiting for public transport for most frequent trip (n=33)
- Q5.9: Safety while using public transport for most frequent trip (n=33)
- Q5.10: Personal monthly transport expenditure (n=33)
- Q5.11: Walking time to nearest public transport access point (n=33)
- Q5.12: Frequency using e-hailing (n=33)
- Q5.13: Involvement in road accident in past 12 months (n=33)
- Q5.14: Places usually walked to (all options) (n=33)
- Q8.14: Most important to give people opportunities in life (n=13)
- Q10.7: Sector of employment (n=3)

Further information on these variables is available in section 5.

In the SPSS version of the dataset, all -1 and -3 values have been stipulated as missing.

In Section 15 (self-complete) all questions had a 'Prefer not to answer' response option. Coding of these responses is question specific, and there is no standardised code. These responses have not been stipulated as missing in the SPSS dataset.

3.3 Standard response options

Default coding for questions with standard response options are listed below. However, the user should be guided firstly by the codes provided in the questionnaire and the labelling within the dataset if this differs from the coding detailed below.

3.3.1 Main questionnaire

1	Yes
0	No

Table 5: Default coding for Yes/No questions

1	Yes
0	No
2	Don't know

Table 6: Default coding for Yes/No questions with 'Don't know' option

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

Table 7: Default coding for satisfaction scale questions

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

Table 8: Default coding for agreement scale questions

1	City of Ekurhuleni
2	City of Johannesburg
3	City of Tshwane
4	Emfuleni
5	Lesedi
6	Midvaal
7	Merafong

8	Mogale City
9	Rand West
10	Don't know
11	Other

Table 9: Default coding for Gauteng local municipalities

1	Very safe
2	Fairly safe
3	Neither safe nor unsafe
4	Bit unsafe
5	Very unsafe

Table 10: Default coding for safety scale questions

For multi-select and yes-no variables, the dataset includes a variable for each response option. The standardised coding for these variables is shown in Table 11. For more information on these question types, see section 2.5.

1	Option selected by respondent
0	Option not selected by respondent

Table 11: Default coding for each response in multi-select or yes/no list questions

3.3.2 Self-complete questionnaire component

Two additional sets of standardised codes were used for the self-complete questionnaire component. These are presented in Tables 12 and 13. Multi-select questions in the self-complete questionnaire component are coded in the same way as those in the interviewer-administered component. Please note that most self-complete section variables are not included in the open access dataset, and must be requested separately from DataFirst.

1	Yes
0	No
2	Prefer not to answer

Table 12: Default coding for Yes/No questions with 'Prefer not to answer' option (self-complete section only)

1	1 time
2	2-3 times

3	4 or more times
4	Prefer not to answer

Table 13: Default coding for frequency of abuse questions

4 RECODES AND DERIVED VARIABLES

The dataset includes a number of recodes and derived variables, which have been used by the GCRO for analysis, and may be of use to data users. This section provides detail and coding for all recodes and derived variables included in the dataset. In all instances the original variable is included as well, should data users prefer to use the original variable or to generate their own recodes. There is one exception to this, which is the age variable. Due to the need to preserve anonymity, we have not been able to include the original age variable, but do provide a version top-coded to 80, along with a fairly fine-grained recode, described in section 4.1.11.

4.1 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. Most recodes contain 'recode' in the variable name, and details are provided below.

4.1.1 Interview duration

The variable 'dur_mins_recode' provides a categorical recode for interview durations in minutes, derived from 'dur_mins'. Categories are provided in the table below.

Value	Label
1	15-20 minutes
2	21-30 minutes
3	31-40 minutes
4	41-50 minutes
5	51-60 minutes
6	61-90 minutes
7	More than 90 minutes

Table 14: Coding of 'dur_mins_recode' variable

4.1.2 Month and year of interview

The variable 'Date_month' is a categorical variable providing the month and year in which the interview was conducted, derived from 'interview_date'.

Value	Label
1	Oct 2020
2	Nov 2020
3	Dec 2020
4	Jan 2021
5	Feb 2021
6	Mar 2021
7	Apr 2021
8	May 2021

Table 15: Coding of 'Date_month' variable

4.1.3 Dwelling type (A3)

The variable 'a3_dwelling_type_recode' simplifies the many categories in 'a3_dwelling_type' into three main categories - 'Formal', 'Informal' and 'Other'. Due to COVID-19 restrictions, the data collection team was not able to gain access to residents of many hostels. For this reason, the proportion of interviews classed as 'other' in QoL 6 is lower than would otherwise be expected.

Original: a3_dwelling_type		a3_dwelling_type_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	Semi-detached house not in a complex	1	Formal
7	House, flat or room separate from main dwelling in backyard	1	Formal
8	Informal dwelling or shack in backyard	2	Informal
9	Informal dwelling NOT in backyard, e.g. in informal squatter settlement or on a farm	2	Informal

10	Room or flat which is part of main dwelling or property	1	Formal
11	Caravan or tent	3	Other
12	Unit in a retirement home or barracks etc	1	Formal
13	Hostel	3	Other
14	Other (specify)	3	Other

Table 16: Recoding of 'a3_dwelling_type' into 'a3_dwelling_type_recode'

4.1.4 Place of birth (Q3.1)

The variables 'q3_1_birth_prov_recode' simplifies the responses in 'q3_1_birth_prov' into three categories - 'Born in Gauteng', 'Born in another province in South Africa' and 'Born in another country'.

Original: q3_1_birth_prov		**q3_1_birth_prov_recode**	
Value	**Label**	**Value**	**Label**
1	Gauteng	1	Born in Gauteng
2	Eastern Cape	2	Born in another province in South Africa
3	Free State	2	Born in another province in South Africa
4	KwaZulu-Natal	2	Born in another province in South Africa
5	Limpopo	2	Born in another province in South Africa
6	Mpumalanga	2	Born in another province in South Africa
7	Northern Cape	2	Born in another province in South Africa
8	North West	2	Born in another province in South Africa
9	Western Cape	2	Born in another province in South Africa
10	Country outside of South Africa	3	Born in another country

Table 17: Recoding of 'q3_1_birth_prov' into 'q3_1_birth_prov_recode'

4.1.5 Timing of respondent's move to Gauteng

The variable 'q3_4_year_gp_recode' is a categorical variable indicating the duration of time since the respondent moved to Gauteng (if applicable). This was calculated first by using the year in which the interview was conducted together with the year in which the respondent moved to Gauteng to determine how many years ago the respondent had moved to Gauteng, and then allocating the appropriate code as shown in the table below.

Value	**Label**
1	In the last year

2	2-3 years ago
3	4-5 years ago
4	6-10 years ago
5	More than 10 years ago

Table 18: Year respondent moved to Gauteng coded

4.1.6 Time of departure for most frequent trip

The categorical variable 'q5_5_depart_time_recode' provides departure time for the respondent's most frequent trip in hourly intervals. The variable is calculated from the time of departure as provided in 'q5_5_depart_time'. The coding for the hourly categories is provided below

Value	Label	Value	Label
1	00:00-00:59	2	01:00-01:59
3	02:00-02:59	4	03:00-03:59
5	04:00-04:59	6	05:00-05:59
7	06:00-06:59	8	07:00-07:59
9	08:00-08:59	10	09:00-09:59
11	10:00-10:59	12	11:00-11:59
13	12:00-12:59	14	13:00-13:59
15	14:00-14:59	16	15:00-15:59
17	16:00-16:59	18	17:00-17:59
19	18:00-18:59	20	19:00-19:59
21	20:00-20:59	22	21:00-21:59
23	22:00-22:59	24	23:00-23:59

Table 19: Coding of 'q5_5_depart_time_recode'

4.1.7 Travel duration for most frequent trip (Q5.6)

The categorical variables 'q5_6_time_destination_recode' provides the duration of the respondent's most frequent trip in 15 minute intervals. This variable is calculated based on the travel duration in minutes as specified by the respondent in the variable 'q5_6_time_destination'. Coding is provided in the table below.

Value	Label
1	0 - 15 minutes
2	16 - 30 minutes

3	31 - 45 minutes
4	46 - 60 minutes
5	61 - 75 minutes
6	75 - 90 minutes
7	More than 90 minutes

Table 20: Coding of 'q5_6_time_destination_recode'

4.1.8 Walking time to nearest public transport (Q5.11)

The categorical variables 'q5_11_pub_transport_prox_recode' provides the walking time to the respondent's nearest public transport access point in 10 minute intervals. This variable is calculated based on the walking time in minutes as specified by the respondent in the variable 'q5_11_pub_transport_prox'. Coding is provided in the table below.

Value	Label
1	0 - 10 minutes
2	11 - 20 minutes
3	21 - 30 minutes
4	31 - 40 minutes
5	More than 40 minutes

Table 21: Coding of 'q5_11_pub_transport_prox_recode'

4.1.9 Not enough money to feed children in household (Q6.5)

The variable 'q6_5_feed_children' has three response options – 'yes', 'no', and 'there are no children in this household'. In some instances, a binary variable offering only 'yes' and 'no' is easier to work with. For this reason, we include the variable 'q6_5_feed_children_binary', coded in this way, with the dataset. Analysts should consider which variable is more appropriate for their purposes, and ensure that they are interpreting the variable accurately. Details are provided in the table below.

Original: q6_5_feed_children		Q6_5_feed_children	
Value	**Label**	**Value**	**Label**
0	No	0	No
1	Yes	1	Yes
2	There are no children in this household	0	No

Table 22: Coding of 'q5_11_pub_transport_prox_recode'

4.1.10 Highest level of education attained (Q14.1)

The categorical variable 'q14_1_education_recode' reduces the responses regarding highest level of education attained, as captured in 'q14_1_education', into six categories. Coding is provided in the table below.

Original: q14_1_education		q14_1_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	2	Primary only
6	Grade 4, Std 2	2	Primary only
7	Grade 5, Std 3 or ABET 1	2	Primary only
8	Grade 6, Std 4 or ABET 2	2	Primary only
9	Grade 7, Std 5	2	Primary only
10	Grade 8, Std 6, Form I or ABET 3	3	Secondary incomplete
11	Grade 9, Std 7, Form II, NQF 1 or ABET 4	3	Secondary incomplete
12	Grade 10, Std 8, Form III, National Trade Certificate 1	3	Secondary incomplete
13	Grade 11, Std 9 or Form IV	3	Secondary incomplete
14	Grade 12, Std 10, Matric	4	Matric
16	A certificate from a college, technikon or university	5	More
17	A diploma from a college, technikon or university	5	More
18	Technikon or university degree	5	More
19	Postgraduate degree - e.g. Hons, MA, PhD	5	More
20	Unspecified	6	Unspecified

Table 23: Recoding of 'q14_1_education' into 'q14_1_education_recode'

4.1.11 Respondent age (Q14.2)

As previously indicated, the raw respondent age variable, 'q14_2_age', is not included in the publicly released dataset, due to the need to protect respondent anonymity. We do, however, include a version of the age variable, 'q14_2_age_topcode', which provides exact age (in years) for all respondents aged up to 79, and then codes all older respondents as 80. In addition, we provide a categorical variable, 'q14_2_age_recode' which converts the original age variable into 11 categories, as detailed in the table below.

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

Table 24: Coding of 'q14_2_age_recode'

4.1.12 Number of household residents (Q14.5)

The variable 'q14_5_people_recode' provides a seven-category alternative to 'q14_5_people'. The recode retains the exact number of household residents for households with up to 6 residents, and collapses all larger households into a single '7+' category, as shown in the table below.

Value	Label
1	1
2	2
3	3
4	4
5	5
6	6
7	7+

Table 25: Coding of 'q14_5_people_recode'

4.1.13 Number of household residents under 18 years of age (Q14.6)

The variable 'q14_6_under18_recode' provides a five category alternative to 'q14_6_under18'. The recode retains the exact number of household residents under 18 for households with up to 3 residents under 18, and collapses all households with larger numbers of residents under 18 into a single '4+' category, as shown in the table below.

Value	Label
0	0
1	1
2	2
3	3
4	4+

Table 26: Coding of 'q14_6_under18_recode'

4.1.14 Number of household residents aged 60 years or older (Q14.7)

The variable 'q14_7_60plus_recode' provides a five category alternative to 'q14_7_60plus'. The recode retains the exact number of household residents aged 60 years or older for households with up to three of these residents, and collapses all households with larger numbers of residents 60 or above into a single '4+' category, as shown in the table below.

Value	Label
0	0
1	1
2	2
3	3
4	4+

Table 27: Coding of 'q14_7_60plus_recode'

4.1.15 Self-reported household income (Q15.3)

The categorical variable 'q15_3_income_recode' reduces the responses regarding monthly household income, as captured in 'q15_3_income', into seven categories. Coding is provided in the table below. Please note that a number of respondents (n=1 850) were not asked this question as they did not consent to participate in the self-completed module. These individuals are coded as '-1' (missing due to skip) and are set to missing in both the original variable and the recode. A further 2 066 individuals selected responses of 'no income', 'don't know', or 'prefer not to answer'. In the original variable, these responses each have a unique code and are not set as missing. In the recode, all are coded as 7, and set to missing. The reason for including those who selected 'no income' in this category was that in a large proportion of cases, other available data from the same respondents suggested that this was not a reliable response.

q15_3income		q15_3_income_recode	
Value	Label	Value	Label
1	R1 - R400	1	R1 - R800
2	R401 - R800	1	R1 - R800

3	R801 - R1 600	2	R801 - R3 200
4	R1 601 - R3 200	2	R801 - R3 200
5	R3 201 - R6 400	3	R3 201 - R12 800
6	R6 401 - R12 800	3	R3 201 - R12 800
7	R12 801 - R19 200	4	R12 801 - R25 600
8	R19 201 - R 25 600	4	R12 801 - R25 600
9	R25 601 - R38 400	5	R25 601 - R51 200
10	R38 401 - R51 200	5	R25 601 - R51 200
11	R51 201 - R76 800	6	R51 201 and more
12	R76 801 - R102 400	6	R51 201 and more
13	R102 401 -R153 600	6	R51 201 and more
14	R153 601 - R204 800	6	R51 201 and more
15	R204 801 - R500 000	6	R51 201 and more
16	More	6	R51 201 and more
17	No income	7	No income/Prefer not to answer/Don't know
18	Prefer not answer	7	No income/Prefer not to answer/Don't know
19	Don't know	7	No income/Prefer not to answer/Don't know

Table 28: Recoding of 'q15_3_income' into 'q15_3_income_recode'

4.2 Derived variables included with the dataset

4.2.1 Access to any form of electricity (Q1.12)

Question 1.12, which asks respondents about all types of electricity supply that they use, was asked as a yes-no list (see section 3.3.1). This results in a series of nine variables, each one representing a particular type of electricity supply, and coded 1 if the respondent said 'yes', and 0 if 'no'. A subset of the variables (q1_12_postpaid, q1_12_8_dont_know and q 1_12_9_none) also have values missing due to skips, as detailed in section 2.8.1.

We use the data from these nine variables to generate the variable 'Any_electricity', coded 1 if the respondent has access to electricity, and 0 if not. We categorised respondents as having access to electricity if they had indicated that they used any of the forms of electricity supply asked about. If they did not report using any of these forms of electricity supply, they were categorised as not having access to electricity. Details are provided in the table below.

Value	Label	Definition

0	No form of electricity	q1_12_1_prepaid=1 OR q1_12_2_postpaid=1 OR q1_12_3_solar=1 OR q1_12_4_generator=1 OR q1_12_5_neighbour=1 OR q1_12_6_car=1 OR q1_12_7_elsewhere=1
1	Some form of electricity	All other respondents

Table 29: Coding of 'Any_electricity'

We note that there are several different ways an analyst might approach creation of this variable, and encourage analysts to assess whether our derived variable is appropriate for their intended use. In particular, some analysts may prefer to make use of the variable q1_12_9_none, in order to rather identify only those respondents who stated that they did not make use of electricity at all. Please see additional information on response patterns to this questions, which may inform analytical decisions, in section 5.1.

4.2.2 Metered electricity connection (Q1.12)

We also used the data from question 1.12 to generate a variable which indicates whether or not the respondent reported using a metered electricity supply, whether prepaid or postpaid, 'Metered_connection'. The coding is detailed in the table below. Please note that in this variable we make use of a 'not applicable/unknown' category for those respondents who report not knowing the types of electricity supply used as well as those who report no access to electricity. In the SPSS version, the 'not applicable/unknown' category is stipulated as missing. Again, we encourage analysts to assess whether this approach is best suited for their particular purposes.

Value	Label	Definition
0	Not a metered connection	All respondents not falling into the two categories below.
1	Has a metered connection	q1_12_1_prepaid=1 OR q1_12_2_postpaid=1
2	Not applicable/unknown	q1_12_8_dont_know=1 OR q1_12_9_none=1

Table 30: Coding of 'Metered_connection'

4.2.3 Currently generating own electricity (Q1.12)

A third derived variable drawing on question 1.12 is 'Generating_electricity', which indicated whether the respondent is already generating energy through either solar, wind or PV power, or through petrol or diesel generators. Coding is shown in the table below. For this variable, we also make use of a 'Not applicable/unknown' category for those who are not sure of electricity source or do not have access to electricity. In the SPSS version, the 'not applicable/unknown' category is stipulated as missing.

Value	Label	Definition
0	Not generating electricity	All respondents not falling into the two categories below.

1	Generating some electricity	q1_12_3_solar=1 OR q1_12_4_generator=1
2	Not applicable/unknown	q1_12_8_dont_know=1 OR q1_12_9_none=1

Table 31: Coding of 'Generating_electricity'

4.2.4 Planning to generate additional energy in the coming year (Q1.12 & Q1.13)

Respondents were asked a series of questions about their plans to generate some or all of their own energy in the following 12 months. Respondents were asked different questions depending on whether they reported in Q1.12 that they were already generating energy through either solar, wind or PV power, or through petrol or diesel generators. As a result, the data from the relevant questions can be somewhat challenging to work with.

We include with the dataset the variable 'Generate_own_electricity', which is coded 1 if the respondent reported planning to start generating any energy in the next 12 months, or increase the amount of energy they were already generating. It is coded 0 if the respondent did not report planning to start generating energy, or increase energy generation, in the next 12 months. Coding is provided below. Please note that individuals who reported in Q1_12_9_none that they have no access to electricity were not asked questions about energy generation. While we have chosen to code them as not planning to start generating energy, analysts may wish to consider alternatives.

Value	Label	Definition
0	Do not plan to generate own electricity	All respondents not meeting the criteria below
1	Plan to generate own electricity	q1_12b_elec_gen_plans=1 or q1_13_elec_loadshedding=1

Table 32: Coding of 'Generate_own_electricity'

4.2.5 Planning to generate all energy in the coming year (Q1.12 & Q1.13)

The dataset also includes a second variable drawing on this set of questions, which indicates whether respondents are planning to start to generate most or all of their own energy (i.e. 'go off the grid' to some extent) in the coming 12 months. This variables is 'Planning_off_grid', and is coded 1 if respondents are planning to do so, 2 if respondents are not sure or answered 'not applicable' due to not owning the property, and 0 otherwise. Coding is provided in the table below. Please be aware that individuals who already generate most or all of their own energy are coded as 0 in this variable, as it particularly focuses on future plans. These individuals can easily be identified by the value of 1 in 'q1_12a_elec_generation'.

Value	Label	Definition
0	Do not plan to start generating most energy in the next 12 months	All respondents not meeting the criteria below
1	Plan to start generating most energy in the next 12 months	q1_12b_elec_gen_plans=1 OR q1_13a_elec_plans=1

2	Unsure/Not applicable	q1_12b_elec_gen_plans=2 OR q1_12b_elec_gen_plans=3

Table 33: Coding of 'Planning_off_grid'

4.2.6 Purpose of most frequent trip (Q5.1 & Q5.3)

Respondents were asked about the purpose of the most frequent trip they make from home in Q5.1. These responses are recorded in the variable 'q5_1_frequent_trip'. However, as detailed in section 2.8.1, those who indicated that they never go anywhere in this question were asked in Q5.2 to verify this. A small number of respondents (n=10) indicated at this point that they did go somewhere, and the purpose of their trip is recorded in 'q5_3_trip'. For ease of analysis, we have created the variable 'q5_1_3_frequent_trip_cons', which uses 'q5_1_frequent_trip' as a base, and then updates the information for the 10 respondents who provided a trip purpose in 'q5_3_trip'. **We recommend using 'q5_1_3_frequent_trip_cons' for any analysis of these questions**.

4.2.7 Most frequently consumed proteins (Q6.6)

Please note that this question was asked as a multi-select question. We provide a dichotomous variable for each protein type, coded 1 if this was selected, and 0 if not. However, we also include a variable ('q6_6_food') which holds the concatenated text names of the two options selected by the respondent. This is a string variable. This variable may be useful if analysts wish to explore the frequency of particular combinations of options.

4.2.8 Participation in organised social activities (Q12.1)

Q12.1, about whether the respondent participated in any of a range of organised social activities, was asked as a Yes/No list. Consequently, the dataset includes a binary variable for each activity type, coded '1' if the respondent reported participating in the activity, and '0' if they did not. The variable 'Social_participation' combines the data from these six variables into a single indicator of whether the respondent participated in any of these forms of social activity in the past year.

Value	Label	Definition
0	No participation	All respondents not meeting the criteria below
1	Some participation	q12_1_1_church = 1 OR q12_1_2_social = 1 OR q12_1_3_stokvel = 1 OR q12_1_4_community = 1 OR q12_1_5_political = 1 OR q12_1_6_other = 1

Table 34: Coding of 'Social_participation'

4.2.9 Involvement in participatory governance activities (Q12.2)

Q12.2, about whether the respondent or a household member participated in any of a range of participatory governance activities, was asked as a Yes/No list. Consequently, the dataset includes a binary variable for

each activity type, coded '1' if the respondent reported participation in the activity, and '0' if they did not. The variable 'Political_participation' combines the data from these seven variables into a single indicator of whether the respondent or other household member participated in any form of participatory governance activity in the past year.

Value	Label	Definition
0	No participation	All respondents not meeting the criteria below
1	Some participation	q12_2_1_ward = 1 OR q12_2_2_street = 1 OR q12_2_3_cdf = 1 OR q12_2_4_idp = 1 OR q12_2_5_mayor = 1 OR q12_2_6_sgb = 1 OR q12_2_7_cpf = 1

Table 35: Coding of 'Political_participation'

4.2.10 PHQ-2 score (Q13.9 & Q13.10)

The QoL 2020/21 included the two four-point scale items that comprise the Patient Health Questionnaire – 2 (PHQ-2), a short screening tool for depressive symptoms (Kroenke et al, 2003). The two items are the frequency of loss of interest or pleasure in things over the past two weeks ('q13_9_pleasure'), and frequency of feeling depressed over the past two weeks ('q13_10_depressed'). Each is coded on a scale running from 'Not at all' (coded 1) to 'Nearly every day' (coded 4)

In order to calculate the PHQ-2, the scores of these two variables are each rescaled to run from 0 ('Not at all') to 3 ('Nearly every day'). The two scores are then added, which provides a single score running from 0 through to 6, with lower scores indicating that the respondent is less likely to be at risk of depression, while higher scores indicate the respondent is more likely to be at risk of depression. This score is available in the variable 'PHQ2_score'. Analysts are advised to consult the relevant literature for guidance on the use of this score.

4.2.11 PHQ2_score_high (Q13.9 & Q13.10)

A widely used approach to the interpretation of PHQ-2 scores is to use a score of three or higher as an indication that an individual is at high risk of depression. The variable 'PHQ2_score_high' is calculated on this basis, and is coded 1 if a respondent is at high risk of depression, and 0 otherwise. Coding is detailed in the table below.

Value	Label	Definition
0	Not at high risk of depression	PHQ2_score < 3
1	High risk of depression	PHQ2_score >= 3

Table 36: Coding of 'PHQ2_score_high'

4.2.12 QoL index variables

The QoL 2020/21 dataset includes key variables related to the GCRO's QoL Index. This section provides only a brief overview of the variables included in the dataset. For details on the derivation and calculation of the Index, please consult Naidoo & de Kadt (2021).

Dimension scores

The dataset includes the score for each of the seven dimensions feeding into the composite QoL Index. These variables are each scaled to run from 0 to 10, with lower scores indicating lower levels of wellbeing in that domain, and higher scores indicating higher levels of wellbeing. Each variable is suitable for use on its own as a measure of wellbeing in that respective dimension. The dimension score variables are as follows:

- Services ('F1servic')
- Socio-economic status ('F2soclas')
- Government satisfaction ('F3govsat')
- Life satisfaction ('F4lifsat')
- Health ('F5health')
- Safety ('F6safety')
- Participation ('F7partic')

Composite QoL Index Score

The overall GCRO QoL Index score for each respondent is available in 'QoLIndex_Data_Driven'. This variable is scaled to run from 0 to 100, with lower scores indicative of poorer quality of life, and higher scores indicative of higher quality of life.

5 ADDITIONAL INFORMATION ON SELECTED VARIABLES AND SPATIAL AREAS

This section provides some additional information on a number of questions and variables, which may be useful to a data user. This covers some implementation challenges impacting a small number of variables, as well as other information relevant to the interpretation of certain variables. This information is drawn from the pilot process, the main data collection process, various debrief activities, and the analysis already undertaken by the GCRO.

Where implementation issues have resulted in missing information or data that is challenging to interpret, we have not attempted to correct the data, unless otherwise specified, preferring to allow each analyst to make their own decision. However, in some instances we have provided recodes in addition to the original data. Where recodes are available, this is indicated, and further information is available in section 4.1

Some information on areas in which achieved sample distribution is skewed is also included.

5.1 Types of electricity used (Q1.12)

Some unexpected response patterns to question 1.12 regarding types of electricity supply makes interpretation of findings slightly complex. While most response patterns across the nine items in this yes-no list make sense, there are 330 respondents (2.4% of the sample) who answered 'no' to all nine options, meaning that we do not actually know whether or not they have electricity – they have answered 'no' to all

possible electricity sources, including 'other', but have also answered 'no' when asked if they have no electricity at all. Discussions with fieldworkers regarding this patterns suggested that some respondents were extremely unwilling to provide any information regarding electricity supply or use. This may relate to use of illegal connections, and concern about possible reprisals.

A further 206 respondents indicated that they don't know what type of electricity supply the household has. While in some instances this is likely to be an accurate response, there is also a possibility that it was used in some cases due to reluctance to disclose illegal connections.

5.2 Electricity generation (Q1.12 & Q1.13)

Due to the complexity of the skip patterns governing which of the electricity generation questions were asked of which respondents, analysis of these variables can be fairly complex. Details of the skip patterns are provided in section 2.8, and details of derived variables which may facilitate analysis are provided in section 4.2. This information should inform analytical decisions in relation to these variables.

Additionally, piloting of these questions, and data collection itself, yielded some information which may further assist in interpretation of the data. These are documented below.

'Going off the grid' and generating electricity for daily use

The term 'going off the grid', which is often used to describe disconnecting from external electricity supplies, was not well understood by respondents during the piloting phase. Consequently, the wording was dropped from the questions, which did introduce challenges in establishing whether respondents were already, or planning to be, completely disconnected from external electricity supplies.

Further, due to the costs of complete disconnection, it was anticipated that too few respondents would be or plan to be completely disconnected. For these reasons, we decided to rather ask about whether energy generation (current or planned) was purely for use during power outages or loadshedding, or whether it was to generate electricity for daily use.

These definitional issues should be considered in analysing and interpreting the data for this set of questions.

Individuals without access to electricity

We made the decision not to ask individuals who reported not having access to electricity any of the questions about electricity generation. Analysts are therefore advised to consider how these individuals should be treated during analysis, as it is not impossible that some may have had plans for electricity generation.

Individuals in accommodation they don't own

Many individuals living in accommodation that they don't own, and also some individuals living in accommodation owned by other family members, struggled to answer the questions about electricity generation. This was either because they had little or no knowledge about plans, or because they had no

authority or ability to make plans. For this reason, most questions included 'Don't know', 'Not applicable' or 'Other' response options. Use of 'Not applicable' was encouraged for respondents in rental accommodation, employer provided accommodation, or hostels, while 'Don't know' was encouraged for use where the respondent was unaware of household plans. Appropriate use of responses in these categories will depend on the nature of analysis being undertaken.

5.3 Country of birth (Q3.2)

Respondents born outside of South Africa were asked for their country of birth in Q3.2. During the course of data collection, however, we became aware that the drop down list of countries available to for use in this question included South Africa, and that 176 respondents recorded in Q3.1 as born outside of South Africa had made use of this option.

Discussions with fieldworkers suggest that many of these respondents were individuals who were reluctant to give further detail on their country of birth, usually due to concerns about whether providing this information might place them at risk. It is likely that in some instances the selection of 'South Africa' was simply a fieldwork error, or the result of an error in capturing the response to Q3.1.

For some respondents, details in interview notes enabled the selection of the correct country in Q3.2, or correction of response in Q3.1. Call backs were made to all other affected respondents to obtain the correct information, and to reassure them that the information would not be shared in ways which might expose them to any risk. This enabled the correction of responses to Q3.1 or Q3.2 for most respondents.

However, there remains 40 respondents who we were not able to reach, or who remained unwilling to disclose their country of origin – this represents 3% of the respondents recorded as born outside of South Africa. As most respondents who were reached did confirm that they were born outside of South Africa, we have left the data for these 40 respondents in 'q3_1_birth_prov' unchanged, reflecting that they were born outside of South Africa. However, we have set the data for 'q3_2_birth_country' to '-3' to indicate that the information is missing due to fieldwork error.

5.4 Section 5 – most frequent trip and transportation questions

During data collection, it became apparent that one of the most frequent uses of the 'Other (specify)' category in Questions 5.1 and 5.3 (purpose of most frequent trip) was for trips made to visit friends and family. This option was added as an additional response option to Q5.1 during the course of data collection, without appropriate testing and validation of skip patterns. Unfortunately, there was a skip pattern error, which resulted in respondents who selected the 'visit friends and family' option skipping all remaining questions in the section. This issue affected 169 respondents.

Callbacks were conducted to address this missing data, and data was obtained from the 136 respondents who could be reached. These records were updated in the database. However, 33 individuals could not be reached, meaning that their records still reflect missing data for all questions from Q5.4 through to Q5.14. As detailed in section 3.2, the missing data is coded '-3', as it is missing due to a fieldwork error. We encourage analysts to consider whether and how the data missing due to fieldwork error might affect their analysis, and adjust their interpretations accordingly.

Please note that in addition, for individuals who did not report going anywhere in Q5.1 and Q5.3 (n=279), the variables Q5.4-Q5.12 and Q5.14 are coded '-1' (missing due to skip), as these individuals were (appropriately) not asked these questions. Further information is available in section 2.8.1

5.5 Satisfaction with food (Q6.7)

Preliminary analysis has raised concerns about the responses provided to Q6.7, about respondent's satisfaction with the food their household usually eats. For this question, over 80% of respondents indicated being satisfied or very satisfied with the food eaten. This is very high given the strong evidence for high levels of hunger and limited access to food revealed in other questions. Indeed, even respondents in households reporting hunger often reported that they were satisfied with the food their household usually eats. Reasons for this pattern is unclear, but fieldworkers thought it might relate to feelings of shame. We advise caution in the interpretation of this variable.

5.6 Most important in giving people opportunities in life (Q8.14)

Towards the end of data collection, we became aware that this question was not marked as required in the data collection application. Consequently, responses were not captured for a small number of respondents (n=13). In these cases, the variable is coded '-3' (missing due to fieldwork error).

5.7 Presence of children in household (Q6.5 and Q14.6)

We collected information on the presence of children in the household in two different parts of the questionnaire. In Q6.5, we asked whether there had been insufficient money to feed children in the household in the past 12 months, and included a response option of 'There are no children in this household". In Q14.6 we asked the respondent to provide us with the number of household residents under the age of 18 years.

For a number of respondents, there is an inconsistency between the responses to these two questions. In Q6.5, 4 945 respondents told us that there were no children in the household. Of these, 176 indicated in Q14.6 that the household had 1 or more resident younger than 18. In addition, of the 6 109 respondents who reported no residents under the age of 18, 96 reported difficulty feeding children in Q6.5, and a further 1 244 reported no difficulty feeding children in Q6.5.

It was not possible to check all responses through call backs, but feedback from field notes and fieldworkers, along with a small number of call backs, suggested a number of explanations for various response patterns. These are:

- Where respondents reported no residents under the age of 18, but responded to Q6.5 indicating no difficulty feeding children, we believe the majority of these cases relate to participants not being offered the full set of response options before responding. That is, we believe that many respondents without children immediately said 'no' when asked about difficulty feeding children, and fieldworkers did not check with the respondent as to whether there were any children.
- We also believe that not all respondents understood the term 'children' in Q6.5 to include all individuals under the age of 18. In particular, feedback indicated that some respondents did not understand the term 'children' to include babies, and that others did not understand it to include adolescents, and particularly older adolescents. We believe that this accounts for the majority of the cases in which respondents reported no children in Q6.5, but reported one or more household residents under the age of 18. We note that this may also be the reason for some of the other inconsistencies noted as well.

- Finally, we believe that changes in household configurations over the preceding 12 months are likely to explain a proportion of the inconsistent responses, and particularly those where respondents reported difficult in feeding children, but no household residents under the age of 18. We believe that in many of the cases in which respondents report difficulty feeding children, but no residents under the age of 18, it is likely that children may have moved out of the household in question. In some instances, children might also have passed away, and a small number might have only very recently turned 18.

5.8 Understanding of 'Gauteng City-Region' as a term (Q8.24 and Q8.25)

Following the initial in-field pilot period, a decision was taken to drop these questions from the questionnaire, as it appeared that many respondents were struggling to understand the meaning of Q8.24, about their familiarity with the term 'Gauteng City-Region'. Those who indicated that they were familiar with the term, and were asked in Q8.25 to explain what the term meant to them also appeared to be struggling. Fieldworkers also reported that these questions cause some confusion, and were therefore time consuming to ask. There was evidence that in many instances respondents conflated the term 'Gauteng City-Region' with the 'Gauteng City-Region Observatory' running the survey.

Both question were removed from the survey, but were erroneously re-introduced shortly afterwards. As a result, the large majority of respondents (all except 1 211) were asked Q8.24. All those who reported familiarity in Q8.24 were also asked Q8.25. Where respondents were not asked Q8.24, the variable is coded as '-1' (missing due to skip). Q8.25 is coded '-1' (missing due to skip) for all respondents who were not asked this question, regardless of whether they had been asked Q8.24.

We have included the data for Q8.24 ('q8_24_gcr') and Q8.25 ('q8_25_gcr') in the dataset, but we suggest that interpretation should take account the feedback that many respondents struggled with these questions. We note also that we have not made any attempt to clean or code the responses to Q8.25, and analysts wishing to make use of this data will need to undertake that process for themselves.

5.9 Sector of employment (Q10.7)

Due to a skip pattern error, individuals who responded in Q10.3 that they had been appointed to a new job, but had not yet started working, were erroneously not asked to answer Q10.7 ('q10_7_sector') about whether they worked in the formal or informal sector. This affected 206 respondents. Through a combination of reviewing fieldwork notes, and call backs to participants, it was possible to provide the response to Q10.7 for all but 3 of the respondents. The variable 'q10_7_sector' is coded '-3' (missing due to fieldwork error) for these 3 respondents.

5.10 Medical conditions of household members (Q13.11)

Question 3.11, implemented as a yes/no list, asked respondents whether they or another household member had had a range of medical conditions during the past year. During training, fieldworkers were trained that the condition should have been diagnosed by a medical practitioner for it to be recorded as yes. Particular emphasis was placed on this for the question on COVID-19 ('q13_11_9_covid'), and this should be taken into account in any analysis of this question, given limited access to testing and diagnosis.

It should be noted that some respondents were more aware than others of the medical conditions of household members, and this should also be considered when working with this data. Where a respondent was unsure, or did not want to respond to a particular item, this was recorded as 'No'. Please be aware that these variables should be used with care when attempting to assess levels of particular medical conditions in an area, as they do not provide information on the number of household members experiencing a particular condition. **This information cannot be used to measure disease prevalence**.

5.11 Experiences since March 2020 (Q13.13)

In question 13.13, respondents were asked to indicate whether they had experienced certain things since March 2020. This was implemented as a yes/no list, with each item read to the respondent, who would then answer with yes or no.

For a subset of these items, a 'not applicable' response option was also available to respondents. The 'not applicable' option was available for:

- Whether the respondent had lost a job ('q13_13_1_job') – it was intended for use by those who had not previously been employed;
- Whether the respondent's working hours had been reduced ('q13_13_2_reduced') – intended for use by those who had not been employed;
- Whether the respondent had permanently closed a business ('q13_13_4_business') – intended for use by those who had not previously owned a business; and
- Whether the respondent has spent more time caring for children or family ('q13_13_5_care) – intended for use by those who did not have caregiving responsibilities.

However, exploration of the use of the 'not applicable' option across these questions suggests that use was somewhat sporadic. This is probably because not all fieldworkers were adequately clear on the appropriate use of this option, and may consequently not have recorded responses correctly or made respondents consistently aware of this option.

5.12 Household income (Q15.3)

In all previous iterations of the QoL survey, interviewers have asked respondents to provide their household income as part of the interviewer-administered questionnaire. Typically, around a third of respondents have refused to answer this question due to its sensitivity.

As QoL 2020/21 included a self-completed questionnaire component, we decided to include the income question in this module, to assess whether this might improve response rates. Unfortunately, this meant that the question was not asked of the 1 850 respondents (13.6% of the sample) who did not agree to complete this section. For these individuals, the variable 'q15_3_income' is coded as '-1' (missing due to skip).

Amongst respondents who did complete this module, response rates were higher than in previous survey iterations. Of this group, 10.8% selected the 'prefer not to answer' option (this is 9.3% of the sample as a whole), and a further 5.3% (or 4.6% of the sample as a whole) selected the 'don't know' option.

In addition, a further 166 individuals selected the 'No income' response option. Based on an internal review of these respondents, we do not feel confident for this group as a whole that this response accurately reflects their household income. For this reason, we recommend caution in working with these responses, and in our recode (q15_3_income_recode) we have adjusted these to missing.

Finally, due to the change in how this question was asked, relative to previous QoL surveys, we encourage analysts undertaking longitudinal comparisons to reflect on whether changes in response patterns might impact on their analyses.

5.13 Frequency of experiencing forms of violence (Q15.10, Q15.13, Q15.18)

Please be aware that these variables are not included in the public version of the dataset, but only in the restricted version.

In a very small number of instances, the skip patterns for particular questions in the self-complete module did not work correctly. This may have been a result of respondents changing answers multiple times, or interacting in other unexpected ways with the data collection interface.

Details are as follows:

- Times hit by current/former partner ('q15_10_times_hit'): Of the 247 respondents who were asked this question, responses are only recorded for 245. Missing responses are coded '-1' (missing due to skip).
- Times kicked by current/former partner ('q15_13_times_kicked'): Of the 168 respondents who were asked this question, responses are only recorded for 167. Missing responses are coded '-1' (missing due to skip).
- Times forced to have sex by current/former partner ('q15_18_times_forced'): Of the 83 respondents who were asked this question, responses are only recorded for 82. Missing responses are coded '-1' (missing due to skip).

5.14 Wards with uneven spatial distribution of interviews

In three of the wards covered by the survey, the distribution of interviews is skewed, with a large proportion of completed interviews clustered in a single EA. This represents a deviation from the sample design (Hamann & de Kadt, 2021). The wards where the distribution was particularly uneven are listed in Table 42 below. There are wards where 34% of interview or fewer are also concentrated in one EA, but these are not deemed significant and not listed here.

These patterns occurred when it was impossible to gain access to any other area in the ward, and in some instances was exacerbated by a skewed distribution of residential dwelling units within the ward. Due to the context of data collection, in practical terms this means that a disproportionate number of interviews were done in lower-income areas in these wards. Analysts engaging in any ward-level analysis should exercise caution when interpreting findings from these wards, as they are not likely to be representative of the adult population of these wards. At higher levels of aggregation, this is not of concern, as these interviews form a small proportion of interviews, and skewness is resolved through weighting.

Ward ID	Description of unevenness
74804019	21 interviews are located in one EA (100% of the interviews for the ward)
74801028	19 interviews are located in one EA (95% of the interviews for the ward)

79800103	14 interviews are located in one EA (54% of the interviews for the ward)
74202009	17 interviews are located in one EA (43% of the interviews for the ward)

Table 37: Wards with uneven interview distribution

6 REFERENCES

GeoSpace International. (2021). GCRO Quality of Life Survey 6: Fieldwork report (2020/21). Johannesburg: Gauteng City-Region Observatory. Available at https://gcro.ac.za/research/project/detail/quality-life-survey-vi-202021/

Hamann, C. and de Kadt, J. (2021). GCRO Quality of Life Survey 6: Sample design (2020/21). Johannesburg: Gauteng City-Region Observatory. Available at https://gcro.ac.za/research/project/detail/quality-life-survey-vi-202021/

Kroenke, K., Spitzer, R.L., & Williams, J.B. (2003). The Patient Health Questionnaire-2: Validity of a two-item depression screener. Medical Care, 1284–1292. https://doi.org/10.1097/01.mlr.0000093487.78664.3c

Naidoo, Y., & de Kadt, J. (2021). Quality of Life Survey 6 (2020/21): Quality of Life Index methodology. Johannesburg: Gauteng City-Region Observatory (GCRO). Available at: https://gcro.ac.za/research/project/detail/quality-life-survey-vi-202021/

Neethling, A. (2021). GCRO Quality of Life Survey 6: Weighting report (2020/21). Johannesburg: Gauteng City-Region Observatory. Available at https://gcro.ac.za/research/project/detail/quality-life-survey-vi-202021/

Orkin, M., 2020. Technical Review of the GCRO QoL Surveys: Synthesis Report. Johannesburg: Gauteng City-Region Observatory. Available online at: https://www.gcro.ac.za/research/project/detail/quality-life-survey-10-year-review/

PMBEJD (Pietermaritzburg Economic Justice & Dignity Group). (2020). Household Affordability Index: December 2020. Pietermaritzburg: PMBEJD. Available at https://pmbejd.org.za/index.php/household-affordability-index/

Statistics South Africa (2012). South African Standard Classification of Occupations (SASCO) 2012. Pretoria: Statistics South Africa. Available at: http://www.statssa.gov.za/classifications/codelists/SASCO_2012.pdf