National Reading Survey 2023

Technical Report

12 June 2023

Contents

Introduction

The National Reading Survey (NRS) is part of the National Reading Barometer project commissioned and managed by the Nal'ibali Trust in partnership with the National Library of South Africa, with additional support from the Zenex Foundation, DGMT and the National Education Collaboration Trust.

The National Reading Survey is a nationally representative survey of 4,250 South African adults aged 16 and above. The National Reading Survey describes the reading practices, preferences and contexts of adults, both in terms of reading for themselves and reading with children in their household. In addition to covering information about frequency, depth and types of reading, access to and preferences regarding reading materials, and attitudes/motivations related to reading, the survey has focus areas relating to library use, digital reading, reading with children and reading language preferences.

This report provides the technical background to the survey and covers:

- Survey and Questionnaire design
- Sampling
- Data collection processes
- Data cleaning and reformulations
- Data analysis
- Technical review conducted on data analysis

Reports on the survey findings can be found at www.readingbarometersa.org.
The questionnaire and final dataset are open-source and can be accessed at DataFirst - Home (uct.ac.za).
We welcome engagement about the survey. Contact us on info@readingbarometersa.org if you:

- Wish to use or adapt the questionnaire for other data collection processes
- Have questions about the dataset or analysis or have found any technical problems with the dataset
- Wish to analyse the data and have questions about variable formulations
- Any other questions you may have

Acknowledgements

The National Reading Barometer project was led by the Nal'ibali Trust, in partnership with the National Library of South Africa (NLSA). It was funded by the NLSA, DGMT, the Zenex Foundation and the National Education Collaboration Trust (NECT). Social Surveys Africa implemented the survey and Social Impact Insights led data analysis on behalf of Social Surveys Africa. twenty8zero7 led communications, with branding and design by Polygram, website by Neil Butcher & Associates and video by Another Love Productions.

The project team was led by Katherine Morse and Katie Huston. Instrument design was led by Katherine Morse. Data collection was led by Lebogang Shilakoe (Social Surveys Africa and Social Impact Insights Africa) with support by Kaytan Ewulu, Musa Mhlanga and Siziwe Sangulukani (Social Survey Africa), data analysis design and writing was led by Tara Polzer Ngwato (Social Surveys Africa and Social Impact Insights Africa), statistical analysis was conducted by Kwame Gyekye and Lovemore Sigwadhi with technical peer review by Ling Ting.

This report was written by Tara Polzer Ngwato of Social Impact Insights Africa with Lebogang Shilakoe, Kwame Gyekye and Katherine Morse.

Suggested reference: Polzer Ngwato, T., Shilakoe, L., Gyekye, G., Morse, K. (2023). National Reading Survey 2023 Technical Report. Nal'ibali Trust.

Background and History of NRS as a Longitudinal Survey

The 2023 NRS is part of a longitudinal series of studies about adult reading in South Africa. The first two iterations of the series in 2006 and 2016 were conducted under the auspices of the South African Book Development Council (SABDC), which was the representative body of the South African book sector until it closed in 2020. This left the literacy sector in South Africa without a source of important information about trends in reading over time. Nal'ibali therefore partnered with National Library of South Africa and a number of donors to collaboratively revitalise and redesign the survey.

The 2023 NRS was designed to enable continuity with the NRS (2016) by retaining a nationally representative sample of around 4000 adults ages 16+ and including many of the same themes and question areas. Full continuity was constrained because the NRB team did not have access to the original 2016 questionnaire, so 2023 questions had to be reconstructed based on how 2016 results were framed in the public report.

While the trendline was maintained where possible (as reported on throughout this report), the NRS 2023 also represents a significant redesign of the questionnaire, including defining 'reading' in broader ways, paying more attention to reading with children, digital reading and reading in multiple languages, and focusing more on the reader's full social experience of reading rather than their behaviour as a consumer of books. In this sense, the NRS 2023 is a 'relaunch' of the series. The current design will be repeated in 2026 and 2030 as a contribution to South Africa's overall efforts to improve child literacy and adult reading by 2030.

Survey Objectives

The National Reading Survey 2023 was designed with the following objectives:

- Understand reading cultures: describe the diverse reading cultures and practices of South African adults.
- Understand reading motivation and attitudes: understand why people read and don't read, including with children, and how they feel about reading.
- Understand access: map access and barriers to access to reading materials, by type and language, including digital materials and reading material for children.
- Track change over time: track changes over time since the National Reading Survey 2016 (conducted by the South African Book Development Council (SABDC)), and when the NRS will be run again in 2026 and 2030, to help reading sector stakeholders understand what is shifting and where more focus is needed.
- Inform research, policy and practice: Contribute to and inspire research that increases understanding of reading in South Africa; promote policy shifts that will create a more enabling context for reading; and inform design of campaigns and interventions to promote and strengthen reading.

Study Audiences

The study was conducted to inform:

- Government policymakers and implementers in the Departments of Basic Education; Higher Education and Training; and Sport, Arts & Culture; Public libraries at National, Provincial and Local municipalities and Metro levels.
- Civil society actors (donors, literacy NGOs and researchers) working on child and adult literacy, and any organisations interested in holistic education, social well-being, social cohesion and social development;
- The publishing and reading materials distribution industry;

- Researchers concerns with education but also adult wellbeing and social development / social cohesion more broadly;
- All South Africans personally interested in reading and their own reading choices and options.

Caveats & Limitations

The NRS 2023 has the following design limitations:

- Excludes school-based reading: The study excludes any discussion of teaching and learning reading in schools, teacher training or the literacy curriculum. These are sufficiently covered by many other studies in South Africa.
- Self-reporting bias: In general, the survey relies on self-reporting about adult and child reading behaviour, which can introduce social desirability bias and does not allow for any assessment of reading 'comprehension' or skill, but this is consistent with the NRS (2016) survey approach.
- No test for reading ability: The survey did not test respondents for reading ability. It relied on self-reported reading ability by asking the question 'if you received a letter would you…' and coding respondents who replied they would give it to someone else to read as 'non-readers'.
- Older Children: Findings on reading by older children (age 10-15) is reported by adults and not the children themselves. Teens aged 16 and above were included in the respondent sample and so have reported on their own reading practices.

Survey Design

The NRS 2023 is part of the larger National Reading Barometer Project, which aims to influence the national reading ecosystem based on the following theory of change:

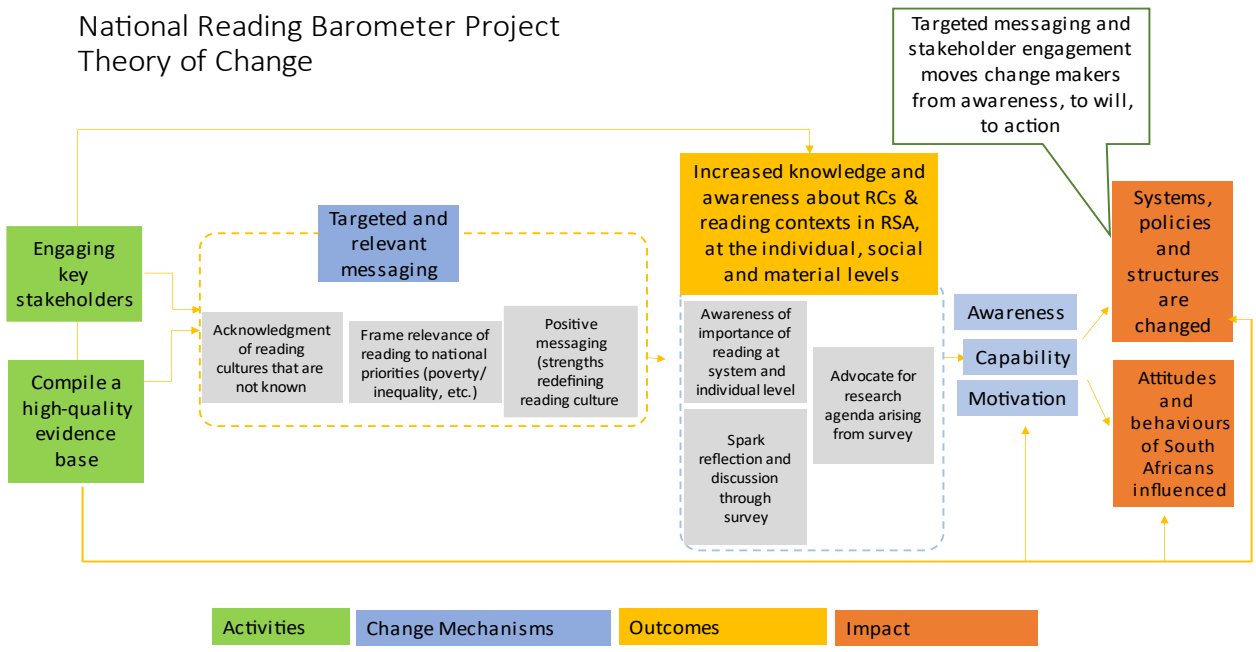


Figure 1: National Reading Barometer Project Theory of Change

The NRS provides the high-quality evidence based from which messaging has been derived (see Summary Report, animated video, infographic and website text at www.readingbarometersa.org) which increases knowledge and awareness about reading cultures & reading contexts in South Africa. This messaging is then used advocate for changes in the reading ecosystem, both at the level of systems, policies and structures, and at the level of attitudes and behaviours.

The intent to contribute as directly as possible to ecosystem change informed the survey question design and especially the analysis process (such as the development of Reader Personas through a statistical clustering process).

Questionnaire Design

The NRS 2023 questionnaire was developed by combining measures reported on in the NRS 2016 report (the original questionnaire was not made public and could not be accessed from SABDC) with new questions derived from the project's conceptualisation of reading, its Theory of Change and consultations with Steering Committee members, a 16-member body representing a wide range of stakeholders in the national reading ecosystem.

The questionnaire was repeatedly revised from June to September 2022. The testing process included Steering Committee members adopting imagined 'reader types' (different ages, backgrounds and reading preferences) and testing the questionnaire from these perspectives. The intent was to formulate the questions in ways that were inclusive of all the different readers and reading experiences in the country. This included a rigorous discussion about any assumptions being made in the question formulation that may have felt alienating or exclusionary to respondents based on culture, socio-economic standing, or language. Inputs received from the Steering Committee resulted in valuable changes to the formulations. The questionnaire was formally piloted in August 2022.

The questionnaire was designed in English. It was not translated in writing, but the field researchers were extensively trained to conduct it in all 11 official languages. This training included identifying key terms and translating these into all languages and carrying out practice interviews in all languages with feedback on translations. Field researchers were allocated to provinces based on the match between the field researchers' own language skills and the dominant languages in that province. All respondents were asked what language they wished to be interviewed in and if the field researcher was not conversant in that language the respondent was called back by another field researcher with the matching language skills.

See section on Lessons Learned for some considerations regarding question formulation and questionnaire design.

Data Collection

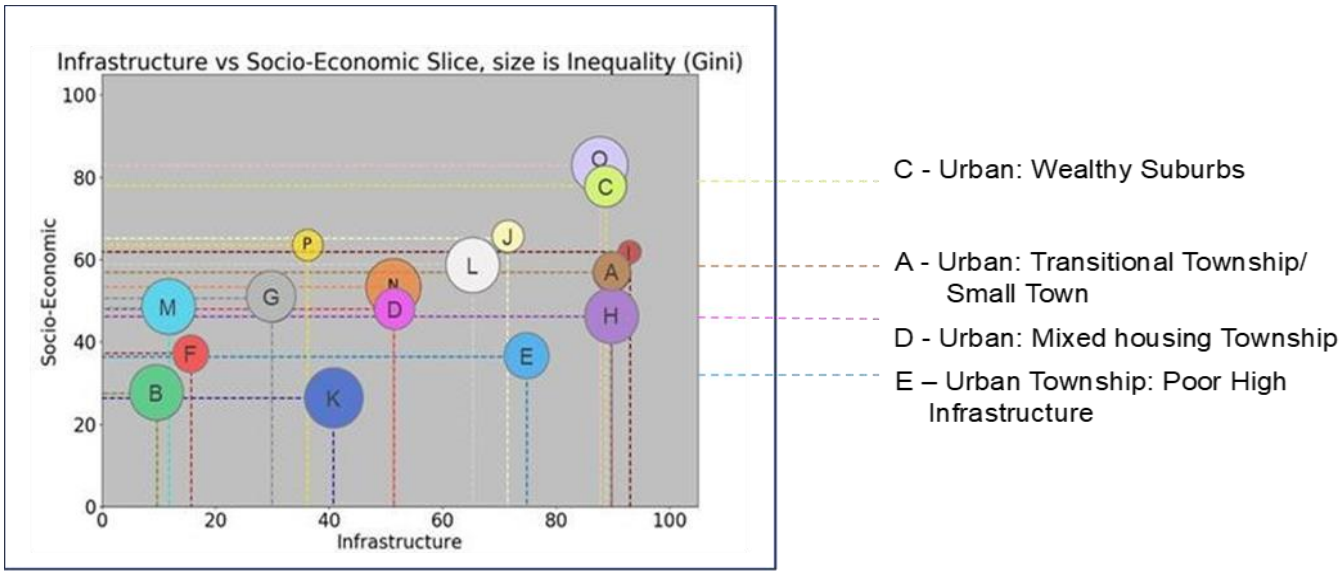
Sampling

The NRS 2023 uses a representative spatial sampling logic. Within a general provincial distribution logic, the sampling logic is based on the recognition that the type of settlement/area in which people live has a significant impact on their opportunities and choices and social and developmental outcomes.

Social Surveys Africa's Community Tapestry© is a unique, rigorous, statistically derived typology that offers interactive, spatially nuanced mapping of community characteristics. It categorises all communities (at 'small area level' (SAL)) in South Africa into 16 main 'types' with shared characteristics based on three dimensions: *Infrastructure, Socio-Economic Standing and Inequality.*

The NRS 2023 used the Community Tapestry to identify neighbourhoods ('communities') to sample within each province. This sample of 'communities' was representative of the socio-economic and infrastructure characteristics of each respective province and therefore of the country as a whole.

Figure 2: Community Tapestry and Cluster type examples



The sampling protocol for the contact harvesters involved allocating a sampled SAL based on their geographic location. Each harvester was provided with the number of households they needed to visit within their allocated SAL, along with a starting location and a sampling interval determined by the number of households in the area. The harvesters followed an alternating sequence, visiting both the main houses and backhouses/ rooms of each household. In cases of vacant households, they would continue to the next household according to the sampling protocol. To ensure a balanced representation, a quota chart was provided, outlining the types of respondents they needed to reach, including different genders and age groups. Harvesters were expected to alternate between these categories during their visits to achieve an even spread across the SAL. Regular reporting, monitoring, and quality control measures were implemented to ensure compliance and address any deviations.

The data collection was conducted in three phases:

1. Contact harvesting phase (November 2022) – Field researchers in all provinces sampled respondents according to the steps listed above, explained the nature of the survey and gained the consent of the respondent to participate in the survey. The respondent's contact details were recorded and they were asked whether they wanted to be interviewed telephonically or if they wanted to complete the survey themselves online. If the respondent chose to self-complete, they were sent an SMS or WhatsApp with a unique identifier code and a link to the online questionnaire. The GPS location, address, age and gender of all sampled respondents were recorded to enable quality control of the sampling process as well as check expected distribution by age and gender. If the respondent had children aged 16-17 in the household, they were asked to provide consent on their behalf and the children's contact details were recorded for telephonic interviewing.
2. Main data collection phase (November 2022 – January 2023, with a break over the year-end holiday period) – Field workers trained in telephonic interviewing used the 'harvested' contacts from phase 1 to telephonically interview respondents who had provided consent and requested a telephonic interview.
3. Top-up data collection phase (February-March 2023) – after achieving the target sample of 4000 respondents in phase 1 and 2, the sample distribution was analysed and gaps identified in the number of White, Coloured and Indian/Asian respondents reached. Using the General Household Survey 2021 as a weighting frame, these gaps were too large to address through weighting (the dataset is weighted by province, population group and age group – see section below on weighting). A top-up sample was calculated targeting the gaps by population groups, province and age group.

Given that it had proven challenging to conduct the contact harvesting process in more well-off suburbs, the top-up sample targeted public areas within the same type of 'community' (i.e. the same type of area, with the same socio-economic and infrastructure characteristics based on the Community Tapestry) where the originally sampled respondents from minority population groups lived. The public areas included shopping centres, gyms and car washes in or near originally sampled suburban neighbourhoods. The top-up sample included respondents from all provinces but with the largest top-ups in Gauteng, Western Cape and KZN.

The 251 respondents interviewed in phase 3 enabled the full combined sample to be reliably weighted to be nationally representative.

Sample Size and Distribution

The NRS 2023 completed a sample of 4251 respondents, aged 16 and above. The original sample size target of 4000 was chosen based on continuity with the 2016 NRS, which also had a sample size of 4000. The sample size enables reliable disaggregation of the sample to provincial level.

The planned distribution by province, based on the overall population distribution is shown in the following graphs and the following table shows the actual vs planned distribution.

Figure 3: Planned NRS 2023 and Actual NRS 2016 Provincial Samples

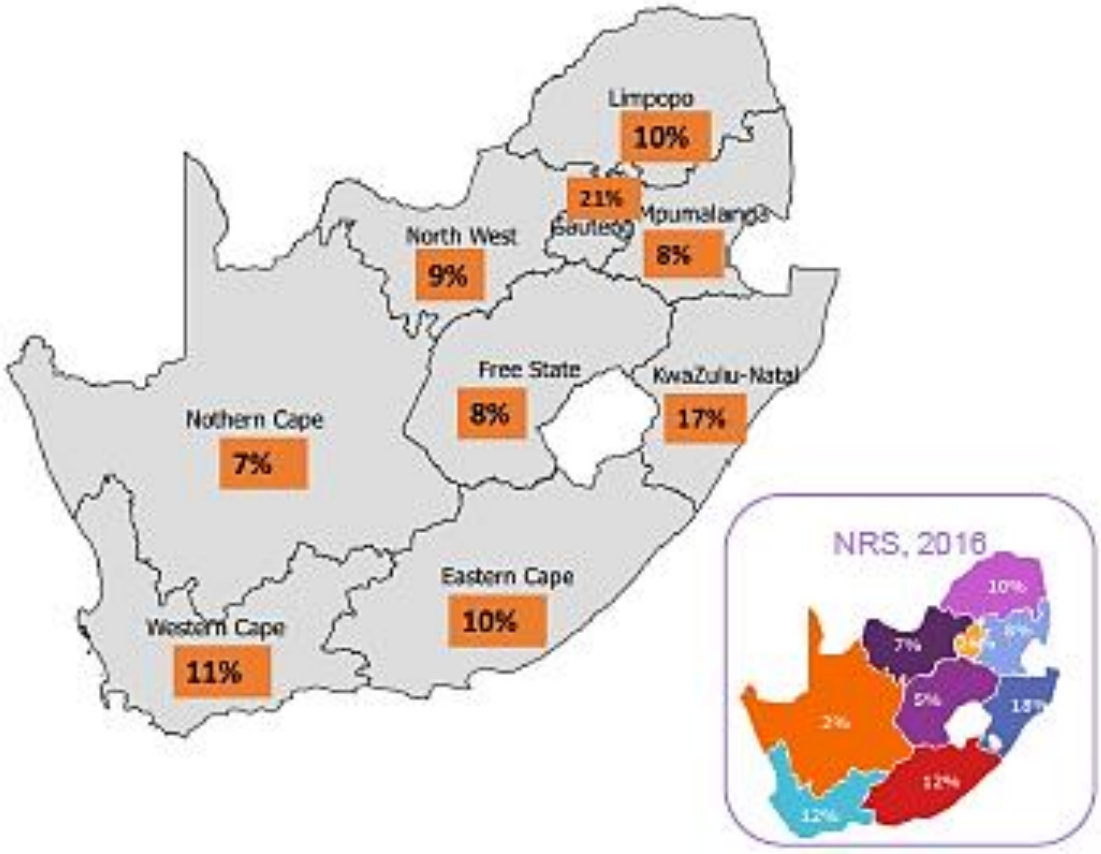


Table 1: Planned vs Actual NRS 2023 Provincial Samples

	planned		actual	
	#	%	#	%
Eastern Cape	417	10%	444	10%
Free State	301	8%	273	6%
Gauteng	822	21%	811	19%
Kwazulu Natal	667	17%	726	17%
Limpopo	388	10%	429	10%
Mpumalanga	310	8%	377	9%
North West	300	8%	464	11%
Northern Cape	355	9%	265	6%
Western Cape	440	11%	462	11%
Totals	4000	100%	4251	100%

Ethics

The survey received ethics clearance from the University of the Witwatersrand Human Research Ethics Committee (Non-Medical), protocol number H22/11/63 in November 2022. The study was rated as low risk.

Data Collection Method

Software

The questionnaire was programmed in Kobo Toolbox.

Data Collector Training

Training of in-field 'contact harvestors' was conducted in August 2022.

Training of field researchers for telephonic interviews was conducted in September 2022 with refresher training after ethics approval was granted in November 2022.

In both cases, training included a strong focus on research ethics and consent, the background and purpose of the survey, detailed engagement with the survey questions (including verbal translation exercises for all key terms and core questions), and practice interviews. All field researchers passed several inter-rater reliability tests before being confirmed into the survey team.

Quality control

Quality data means two things:

- Data is complete: all questions for each respondent are answered
- Data is accurate: what is recorded in the questionnaire is actually what people said.

Data completeness

- Programming checks: the questionnaire programming requires completion of all questions to mitigate against missing data, and automated skip patterns so that you cannot miss out on sections by accident.
- Manual checks: every completed interview is automatically uploaded to the SSA central server as soon as the instrument is closed on the tablet. SSA quality control staff check every uploaded interview as it comes in on the server to see if information is missing so if there are responses which seem too similar or too different from other respondents (especially comparing responses lodged by the same interviewer). Suspicious interviews are included in the call-back roster (see below).

Data accuracy:

- Manual checks: the manual checks mentioned above also check for suspicious repetition of responses, which may indicate that a Field Researcher is entering fake data.
- Check backs: all respondents were asked if they were willing to share their cell phone numbers for quality control purposes. At least one randomly selected respondent per Field Researcher per day was phoned back to ask if the person was indeed interviewed and to check answers to a few of the questions to see if the answers match those originally recorded.
- Time checks: start and end time of the interview is recorded to monitor the time taken to complete the interview. Surveys completed in a shorted than average time were flagged for call backs and audio recording checks.
- Audio recordings: the telephonic surveys were audio recorded and random samples of these recordings were listed to by quality controllers to test for interviewing technique, translation fidelity and accurate capturing of responses.
- We provide space for interviewer comments to enable interviewers to give a summary of his/her observations and report issues that were not covered by survey questions.

Data Cleaning & Reformulation

Data cleaning and reformulation includes the following elements: weighting, treatment of outliers, treatment of missing values due to skip patterns and other reformulations.

Data cleaning and reformulation was carried out in STATA. The anonymised raw data download and STATA cleaning .do file can be shared on request.

Variable renaming

The questionnaire was programmed with variable names that were not self-explanatory (i.e. D4, BM10, etc.). The data cleaning process therefore included renaming all the variables in the dataset (included in STATA .do file). The original questionnaire names are included in the clean dataset as variable labels.

Outliers

The survey only included continuous variables related to a) time spent on various activities and b) amount of money willing to pay for a new book.

Number of hours spent on an activity in a week

Adults:

- reading for enjoyment,
- reading for information,

- reading to communicate,
- on social media,
- watching TV/movies,
- listening to the radio

Older children (10-18 as reported by their caregivers): number of hours spent on an activity in a week:

- reading for enjoyment,
- reading for information,
- reading to communicate

The following rules were applied to outliers for these variables:

- for each variable, placeholder numbers for 'don't know' – 168 which is the total number of hours in a week – were removed and replaced with missing values.
- for each variable, responses above 112 (waking hours in a week) were removed (recoded as missing values). These variables were renamed with the naming convention Cl (cleaned), e.g. hrspw_enjoym_Cl_2_23_1, hrspw_info_read_Cl_2_23_2, hrspw_read_comm_Cl_2_23_3
- for adult reading volumes ONLY (hours per week spent reading for enjoyment, information and communication), a composite variable was constructed called hrspw_total_read_2_23 which adds up the total number of hours per week spent reading for enjoyment, information and communication. Any values above 112 for this composite variable were removed.
- for adult reading volumes ONLY a second version of the clean variable was constructed with the naming convention Cl2 (hrspw_enjoym_Cl2_2_23_1, hrspw_info_read_Cl2_2_23_2, hrspw_read_comm_Cl2_2_23_3.) For these variables, responses to each constituent variable were removed if the hrspw_total_read_2_23 variable for that respondent was above 112. This version of the reading volume variables was used in the calculation of the reading volume averages used in the findings reports.

Note that the variables which coded the continuous variables for hours per week spent on activities into 'buckets' of never, low (1-4 hours per week), medium (5-10 hours), high (11-20 hours) and very high (21 hours and above) (see below in section on reformulation) retained the outlier responses (over 112 but not placeholder numbers) in the category 'very high'.

Amount that would spend on a new book

This variable (buy_bk_pay_how_much_Cl_2_35) was analysed for the findings report by excluding values above R500 (53 responses). The coded version of this variable (buy_bk_pay_how_much_rcd_2_35) includes these responses in the highest 'bucket' ('more than R300).

Skip Patterns

The NRS 2023 questionnaire used skip patterns to reduce the burden of questions that would not apply to a respondent based on their previous answers to related questions. This is standard practice and was extensively piloted to ensure that skip patterns did not incorrectly miss information. A few challenges in the skip patterns nonetheless occurred.

This report section includes clarification on the skip patterns and the effects on the size of various sub-samples (sampled excluding categories of people who were skipped for various reasons). It also summarises how these skips were treated in the cleaned dataset and the analysis.

The Lessons Learned section of the report includes further notes on challenges in the skip patterns and recommendations on changing these for future iterations of the survey or other researchers using the questionnaire in future.

Variables resulting in skips of other questions

Summary of skipping variables and the sub-samples affected:

Skip name	Variable name in clean dataset	Variable question and response based on which respondents were skipped	Number of respondents skipped based on this question and remaining sample
Inability to read	inability_read_letter_1_19	1.19 Imagine that a letter arrives for you from a friend or relative. Which of the following would you do with this letter? Response: I would ask someone to read it to me	429 skipped 3822 remaining sample
Live with children	live_with_children_1_16 (this is a recode of the raw response into a dichotomous yes/no variable)	1.16 How many children under the age of 18 live with you? This can include your children, grandchildren, younger siblings, or children you look after Response: 0	2179 skipped 2071 remaining sample
Ever read with young children	rwc_youngch_everread_cat (this is a recode of rwc_youngch_evread into a dichotomous yes/no variable)	3.2 Do you ever read with the young children you live with? Response: No	459 never read with the young children they live with 1009 remaining sample of caregivers who do read with the children they live with (who responded to this question - see notes below on other missing values for this variable)
Ever access to internet	internet_cl_1_9 (this is a recode of internet_use_1_9 into a dichotomous yes/no variable)	1.9 How frequently do you use the internet? Response: Never	1211 never access the internet 3021 remaining sample
Have & Own no books in the home	bks_num_all_2_24_11	2.24.11 How many books are in your home right now, including books you own and those you have borrowed? Response: None	1080 have no books in the home 3171 remaining sample

Some questions in the Reading with Children section are furthermore skipped based on whether the respondent reported living with young children (in the age categories 0-5 and 6-10) or older children (11-14 and 15-18).

In the dataset, the default coding for questions that were not asked based on skips is 'missing values'. This does not, however, enable a distinction between missing values based on skips and those based on non-response to that question. Furthermore, it does not enable analysis based on skips with substantive information (i.e. a respondent who was not asked how often they use

social media because they never access the internet can be assumed to 'never' use social media) and those where a skip cannot be assumed to provide information on the answer to another question (i.e. responses to the question 'were you ever read to by your parents when you were a child' cannot be assumed for respondents who were skipped based on not living with children today).

For these reasons, the missing values due to skips were treated in the following ways in the dataset and analysis:

- In cases where the results are reported out of a sub-sample (e.g. % of caregivers, or % of internet users), the original variable is used, retaining skipped values as missing values which are not reported on

- In cases where results are reported out of the total population of adults, substantive missing values due to skips have been recoded into new variables (using the naming convention _MV in the clean dataset).

  Codes for missing values are:

  a. inability_read_letter_cat_1_19==0 is recoded as 998
  b. internet_cl_1_9==0 is recoded as 997
  c. rwc_youngch_everread_cat==0 is recoded as 996
  d. bks_num_all_2_24_11==0 is recoded as 995

- In cases where responses are not implied the skip, skipped values are retained as missing values (i.e. responses to the question 'were you ever read to by your parents when you were a child' cannot be assumed for respondents who were skipped based on not living with children today)

Some questions are skipped on multiple prior questions, i.e. questions about reading online materials to children are skipped based on reading ability, internet access and whether the caregiver ever reads with their children.

The STATA files for the analysis show which version of each variable (with or without missing value recoding) was used to generate each finding in the survey findings reports.

The 'relevance' column in the programmed instrument (available on DataFirst) includes all the skip logics.

More detail on the skip patterns is described below.

Ability to Read (inability_read_letter_1_19)

Question: 1.19 Imagine that a letter arrives for you from a friend or relative. Which of the following would you do with this letter?

D20_1	Freq.	Percent	Cum.
I would ask someone to read it to me	429	10.09	10.09
I would read it easily myself	3,540	83.27	93.37
I would read it myself, but it would be difficult	230	5.41	98.78
Refused to answer	52	1.22	100.00
Total	4,251	100.00	

The 429 respondents who answered that 'I would ask someone to read it to me' were considered to be 'unable to read'. Based on this response, therefore, further questions throughout the questionnaire which were concerned with reading practices were skipped. These included:

Variable label	Question
D21	1.20 What languages can you read and write in?
A7_note	How often do you...
A7_1	2.22.1 Read for your own personal enjoyment (like reading for entertainment or relaxation)
A7_2	2.22.2 Read to get information or instructions or learn something (like reading newspapers, for work, for studies, instruction manuals)
A7_3	2.22.3 Read and write to communicate with others (like reading letters, sms, chat messages, social media posts, emails)
A7_note_2_1	Now we have some questions about how much time you spend on different kinds of activities, like reading for pleasure, reading for information, reading for communication, being on social media, watching TV, listening to the radio, etc. You may even do two things at the same time sometimes (like watch TV while being on social media). Please estimate the amount of time you spend on each activity in a week.
A7_1_a	2.23.1 How many hours per week do you generally spend reading for **personal enjoyment** (not work or study)
A7_2_a	2.23.2 How many hours per week do you generally spend reading **to get information or instructions or learn something** (for work, study or because you choose to)
A7_3_a	2.23.3 How many hours per week do you generally spend **reading and writing to communicate with others**
A7_4_a	2.23.4 How many hours per week do you generally spend on **social media**, including WhatsApp, Facebook, TikTok, Instagram, Twitter, etc.?
A29_note	**Reading Material types** I am going to mention several types of things to read. For each, how often do you read these things for yourself (not for children or others)? The answer options are never, a few times a year, about once a month, several times a month, several times a week, daily.
A29_note1	How often do you read ... (for yourself)
A29_A	2.25.1 Print Magazines
A29_B	2.25.2 Print Comic books
A29_C	2.25.3 Print fiction (novels, stories)
A29_E	2.25.4 Print Non-fiction books (informational, documentary)
A29_F	2.25.5 Print Religious books
A29_G	2.25.6 Print Newspapers
A29_H	2.25.7 Magazines, blogs, opinion pieces online and on mobile apps
A29_I	2.25.8 Fiction (novels, stories) online and on mobile apps
A29_J	2.25.9 Non-fiction books online and on mobile apps
A29_K	2.25.10 Religious texts online and on mobile apps
A29_L	2.25.11 News online and on mobile apps
A29_M	2.25.12 Social media posts
A29_N	2.25.13 Downloaded e-books or on mobile apps
A26	2.26 Which of the following statements best describes your preferred format for reading books (on any topic)?
BM13_note	**Online communications** The next few questions are about online communications
BM13_start	How often do you do the following?

BM13	How often do you...
BM13_1	2.27.1 Read and write emails
BM13_2	2.27.2 Type and read chat messages (e.g. SMS, WhatsApp, Messenger)
BM13_3	2.27.3 Search for information online (using Google or any other internet search engine)
BM12_2	2.28.2 I read only if I have to.
BM12_4	2.28.4 I like talking about books with other people.
BM12_5	2.28.5 For me, reading is a waste of time.
BM12_6	2.28.6 For me, reading is a way to explore new people and places.
BM12_7	2.28.7 I read only to get information that I need.
BM12_8	2.28.8 I read as a way to improve my economic situation.
BM12_9	2.28.9 If people in my family saw me reading a book, they would make fun of me.
BM12_10	2.28.10 If my friends saw me reading a book, they would make fun of me.
BM12_11	2.28.11 Reading helps me relax.
BM12_12	2.28.12 Reading is stressful for me.
BM20_note	How much do you agree or disagree with the next set of statements about reading (remember this includes all types of reading)
BM20	How much do you agree or disagree with this statement...
BM20_1	2.29.1 I would read more if things to read were more affordable.
BM20_2	2.29.2 I would read more if I could find things to read that are free.
BM20_3	2.29.3 I would read more if I felt more confident as a reader.
BM20_4	2.29.4 I would read more if I could find interesting things to read.
BM20_5	2.29.5 I would read more if my friends and family also read and talked about what they are reading.
BM20_6	2.29.6 I would read more if there was more to read in my preferred languages.
BM20_7	2.29.7 I would read more if the stories, characters or information were more like my daily life.
BM22	2.29.8 How would you describe yourself? Which of these categories do you identify with most:
A23	2.34 If there was a popular new book you wanted to buy, where would you prefer to buy it from?
A30_1	3.2 Do you ever read with the young children you live with?
A30_1a	3.13 Do you ever read to or with the older children?
BM15	3.35 Do you belong to a book or reading club?
A31_note	Reading Initiatives
A31_C	3.36.1 World Book Day
A31_D	3.36.2 Nal'ibali
A31_E	3.36.3 Read to Lead
A31_F	3.36.4 Book Dash
A31_G	3.36.5 African Storybook Project
A31_H	3.36.6 FunDza
A31_A	3.36.7 National Book Week
A31_B	3.36.8 Kha Ri Gude
A3	4.5 How many e-books do you have?
Text_length_note	Now please think about all the different kinds of things you read: for fun, work or communication; in print, online or on mobile apps. How often do you read...
Text_length	How often do you read...
BM41_1	4.9.1 Short messages: a few words or sentences like a twitter post, whatsapp chat, cookbook recipe or prayer
BM41_2	4.9.2 Short articles: a few paragraphs, like a short newspaper article, bible passage or long email
BM41_3	4.9.3 Medium text: a few pages, like a short story or a long newspaper article or blog post, or a section in a religious text, or a textbook chapter
Text_length_end	4.9.4 Long text: many pages like a novel or non-fiction book or a religious book

Live with Children

Question: 1.16 How many children under the age of 18 live with you? This can include your children, grandchildren, younger siblings, or children you look after.

Response: 0

RECODE of no_children_live_with_1_16 (D9)

	Freq.	Percent	Cum.
No	2,179	51.27	51.27
Yes	2,071	48.73	100.00
Total	4,250	100.00	

The 2179 who responded 0 were skipped for the following questions:

D9_2	1.17 How old are the children you live with
Reading_children_note	**Section 3: Children** The next section is about children and reading with children
BM5	3.1 Which of these statements is mostly true for you...
young_child_note_start	Let's talk about young children living with you, who can't yet read or who are learning to read
BM21	3.12 Thinking about the oldest child in your house, how old were they when they got their first children's book?
middle_child_note	Let's talk about the slightly older children in your house who can read for themselves, so children aged 10 and older. If there is more than one older child, think about the one who reads the most.
BM10_child	3.14 How often does the older child read for themselves at home?
A30a	3.15 Does this older child have a phone of their own?
A7_1_a_child	3.16 How many hours per week does the child generally spend reading for personal enjoyment (not school)
A8_device_child	3.17 When reading for personal enjoyment, how often does the child read on a phone?
A7_2_a_child	3.18 How many hours per week does the child generally spend reading to get information or instructions or learn something
A9_device_child	3.19 When reading for information or to learn something, how often does the child read on a phone?
A7_3_a_child	3.20 How many hours per week does the child generally spend reading and writing to communicate with others. This includes reading & writing on social media but not watching videos or looking at pictures on social media
A30_2_child	3.21 Which language(s) do the older children read in?
A30_3_child	3.22 Which language(s) would you like the older children to read in?
A14_1_child	3.23 How many books for older children do you have at home right now? I mean books written for the children's age group that they can read themselves. This includes books you own and those you have borrowed.
BM4_14	3.25.14 I would read more with children if I could find things to read that are free
BM4_15	3.25.15 I would read more with children if I felt more confident as a reader.
BM4_16	3.25.16 I would read more with children if I could find interesting things to read.
BM4_17	3.25.17 I would read more with children if there was more to read in my preferred languages.
BM4_18	3.25.18 I would read more with children if the stories, characters or information were more like my daily life

BM4_19	3.25.19 I would read more with children if I had more time
A44	3.26 What language is your child mainly taught in at school? This is about the language used for teaching subjects (like maths) to your child, not languages taught as a second language.
A45	3.27 Have any of your children's teachers ever communicated with you about reading with children? This could be through a letter, a meeting, etc.
A16_2	3.28 Do the children you live with have access to a school library?
A18_note	Books from school
A18_start	How often do the children you live with bring home the following types of books from school?
A18	How often do the children bring home…
A18_1	3.29.1 Readers that are part of homework
A18_2	3.29.2 Books borrowed to read for fun
A18_3	3.29.3 Text books
BM14	3.34 How many of the children you live with currently belong to a book or reading club?

Do you ever read with young children

Question: 3.2 Do you ever read with the young children you live with?

Response: No

tab rwc_youngch_everread_cat_MV

RECODE of rwc_youngch_evread (A30_1) |

	Freq.	Percent	Cum.
No	459	28.56	28.56
Yes	1,009	62.79	91.35
998	139	8.65	100.00
Total	1,607	100.00	

Additional notes on missing values for rwc_youngch_everread_cat_MV

- 139 of the missing values were recoded as 998 and included in the analysis because they represent caregivers of young children who are not able to read and were therefore skipped on Ability to Read (inability_read_letter_1_19)
- This question, although directed at caregivers of young children, was actually asked of all caregivers and so the total respondents should be 2071. However, only we only have responses for 1607 (including the 139 recoded as 998), meaning there are 464 missing values. For this reason, this variable was not used extensively in the findings reports.

Variable label	Question
A30_2	3.3 Why do you not read with the young children you live with?
BM10	3.4 How often do you read with these young children?
BM11	3.5 What time of day are you most likely to read with these children?
A30_2	3.6 When reading with the children, which language(s) do you read in?
A30_3	3.7 When reading with the children, which language(s) would you prefer to read in?
A45	3.8 When reading with the children, what kinds of things do you read with them?

Never access the internet

Question: 1.9 How frequently do you use the internet?

Response: Never

RECODE of internet_use_1_9 (A1)

	Freq.	Percent	Cum.
Never use internet	1,211	28.62	28.62
Ever use internet	3,021	71.38	100.00
Total	4,232	100.00	

Note there are 19 missing responses for this question.

Questions skipping on this response:

A29_H	2.25.7 Magazines, blogs, opinion pieces online and on mobile apps
A29_I	2.25.8 Fiction (novels, stories) online and on mobile apps
A29_J	2.25.9 Non-fiction books online and on mobile apps
A29_K	2.25.10 Religious texts online and on mobile apps
A29_L	2.25.11 News online and on mobile apps
A29_M	2.25.12 Social media posts
A29_N	2.25.13 Downloaded e-books or on mobile apps
BM13_note	**Online communications** The next few questions are about online communications
BM13_start	How often do you do the following?
BM13	How often do you...
BM13_1	2.27.1 Read and write emails
BM13_2	2.27.2 Type and read chat messages (e.g. SMS, WhatsApp, Messenger)
BM13_3	2.27.3 Search for information online (using Google or any other internet search engine)
A19	3.31 Do you ever access free books or stories online or on mobile apps?
A20_note	Free story sites and mobile apps
A20_start	Which of the following free online story sites or mobile apps are you aware of and use?
A20	Are you aware of and use...
A20_other	3.33 Which other free online story sites or mobile apps are you aware of?
A31_start	Have you heard of or participated in any of these reading initiatives?
A31	Have you heard of or participated in...
A2	4.1 Which device are you most likely to use to access the internet?
A2_1	4.2 Who owns this device that you use most to access the internet?
A4	4.3 On a scale from 1 to 5, where 1 is never available and 5 is always available. How would you describe your access to the internet?
A5_note	Internet Access Locations
A5_start	How often do you access the internet from the following places...
A5	How often do you access the internet from...
A3	4.5 How many e-books do you have?
A3_0	4.6 How many audio books do you have?
social_media	**Social media presence** I will now ask you about your use of social media. Social media includes WhatsApp, Facebook, Youtube, TikTok, Instagram, etc.
BM8_0	4.7 Are you on social media?

Books in the home

Question: 2.24.11 How many books are in your home right now, including books you own and those you have borrowed?

Response: 0

Questions skipping on this response:

A13	2.24.12 In what languages are the books in your home?
A40	2.24.13 Are any of the books in your home written by a South African author?
A27	2.30 Do you buy new or second-hand books?
A28	2.31 Where do you usually get books?
A28_pref	2.32 Where would you prefer to get books?
A14_1	3.9 How many children's picture books do you have at home right now? By picture books I mean books with only pictures or with pictures and words that young children can look at themselves or that adults can read to young children. This includes books you own and those you have borrowed.
A14_2	3.10 How many children's learn-to-read books (or 'readers') do you have at home right now? By learn-to-read books I mean books with simple words and sentences that young children can read themselves or together with adults. This includes books you own and those you have borrowed.
A14_1_child	3.23 How many books for older children do you have at home right now? I mean books written for the children's age group that they can read themselves. This includes books you own and those you have borrowed.

Weighting

The survey was weighted based on the distribution of province, population group (race) and age group. The reference dataset was the StatsSA General Household Survey 2021.

Weights of between .018 and 4.03 were applied, however only 47 respondents (0,01% of the total sample) had weights above 2. This included 41 White respondents (26% of the 154 total White respondents in the sample), reflecting the challenge of successfully reaching White respondents, despite the targeted efforts made, including an additional two weeks of top-up data collection focussed on minority population groups.

Variables Reformulation

Reformulated variables in the clean dataset either have _cl, _rcd or _cat in the variable name.

Types of variable reformulations in the clean dataset include:

Coding quantity responses into buckets:

- Age into age_groups
- Frequency questions were recoded into broader frequency groups:
    a. Frequently = daily or several times a week. For adult reading, frequently includes reading daily and several times a week. For reading with children, more response options were given for 'how often': frequently includes reading daily, almost daily and 2-3 times a week.
    b. Regularly = weekly, several times a month or monthly
    c. Rarely = less than monthly
    d. Never = never
- Hours per week questions were recoded into buckets:
    a. Never = 0
    b. Low = 1-4
    c. Medium = 5-10

d. High = 11-20
e. Very high = 21 and above

The cut-off points between categories are based on an analysis of the distribution of responses.

- Amount that would spend on a new book was recoded into buckets:
  a. Under R20
  b. R20-R50
  c. R51-R100
  d. R101-R150
  e. R151-R200
  f. R201-R250
  g. R251-R300
  h. More than R300

Coding attitude responses for 'direction'

- Attitude and motivation questions were formulated both positively (Reading helps me relax) and negatively (Reading is stressful for me). To enable consistent analysis for whether the respondent had overall positive or negative attitudes toward reading, responses to negatively worded questions were inverted. The original variables remain in the dataset. The inverted variables are named with _rcd_rev including:
  a. motiv_read_havto_rcd_rev_2_28_2
  b. motiv_waste_o_t_rcd_rev_2_28_5
  c. motiv_family_fun_rcd_rev_2_28_9
  d. motiv_frnds_fun_rcd_rev_2_28_10
  e. motiv_stressful_rcd_rev_2_28_12

Simplifying responses

- Employment: the original employment status variable (employment_status_cde_1_10) recoded 'other' and was simplified into
  a. Unemployed = not working - looking for work
  b. Employed = working full time, working part time, unpaid volunteer
  c. Not economically active = retired, not working not looking for work
  d. Student = student

- Education: the original education variable (education_6_1) recoded 'other' and was simplified into:
  a. No schooling = no schooling
  b. Primary = primary
  c. Secondary = incomplete secondary, complete secondary
  d. Tertiary = incomplete further education, complete further education

- Household income: the imputed hh income variable (hhincome – see below on imputation) was simplified into a smaller number of categories. This allows for testing of the point at which income differences become statistically significant in bivariate analyses and regressions:
  a. Hhincome3: R0-R3200; R3201-R12800, R1280a and above
  b. Hhincome2: R0-R12800; R12801 and above

Data Analysis

This section describes the following types of analysis conducted with the data:

- Calculations of average reading volume
- Imputations of household income
- Construction of compositive variables
- Analyses resulting in Reader Personas, including indices for different dimensions of reading and clustering of these indices into 'Personas'.
- Regressions

Measures for Reading Volume (Amount of Time)

The survey findings report includes reported average reading volumes (number of hours per week spent reading for enjoyment, information and communication). After addressing the outliers in the recorded 'hours per week' variables (see above on Outliers), these averages are based on the mode, since the distribution of responses is skewed which makes the mean an inappropriate measure of average distribution. The mode was identified through the STATA command sum var, detail with unweighted data.

Imputation of household income

A question about household monthly income was asked at the end of the survey (6.3 What is your total household monthly income'), with answer options: R0-R1600; R1601 - R3200; R3201 - R6400; R6401 - R12800; R12801 - R25600 ; R25601 and above; Do not know; Refused to answer

590 respondents answered 'don't know', 833 'refused to answer' and 14 were missing values, i.e. a total of 1437 or 34% of the total sample had no valid information about household income (see variable household_income_6_3). We therefore imputed household income for these respondents based on the following variables:

- Employment status
- Education level
- Dwelling type
- Age group
- Gender

All the imputation was done in R using the MICE function in the MICE package. Multivariate Imputations by Chained Equations (MICE) operates under the assumption that given the variables used in the imputation procedure, the missing data are Missing At Random (MAR), which means that the probability that a value is missing depends only on observed values and not on unobserved values. A default number of five multiple imputations was used and the predictive mean matching approach was used for the imputation.

The imputed dataset was then merged into the clean survey dataset as the variable hhincome and this variable was used for all further analysis requiring socio-economic information, including the regressions.

Composite Variables

Composite variables were constructed to aid analysis.

- Language groups: multiple questions were asked about languages, including languages spoken at home, book ownership, actual and preferred reading languages, languages read to children, etc. In each case, all eleven official languages were listed in the questionnaire, either as single choice or mostly as multiple choice. These language responses were then recoded into language groups for ease of analysis. The following recoded group variables were added to the dataset:
    a. Nguni: isiZulu, isiXhosa, isiNdebele, Siswati

b. Sot-Ped-Tsw: Sesotho, Sepedi, Setswana

c. VenTso: Tshivenda, XiTsonga

d. EA: English, Afrikaans

e. Indigenous: all nine official languages apart from English and Afrikaans

f. Multi: if for multiple choice questions on languages the respondent said they speak/read in multiple languages

- Ever read: This variable was constructed for the findings report and the regressions and represents whether the respondent ever reads in any way. The variable combines whether the respondent *can* read (inability_read_letter_cat_1_19 !=0) and whether they *do* read (if all variables for number of hours per week spent reading for enjoyment 9 hrspw_enjoym_Cl_2_23_1), information (hrspw_info_read_Cl_2_23_2) or communication (hrspw_read_comm_Cl_2_23_3) are above 0.
- How frequently read 'books': in the findings report, several findings relate to whether people 'read books'. These are based on a composite variable that combines whether the respondent reads print fiction (readfrq_mat_prntfict_rcd_2_25_3), print nonfiction (readfrq_mat_prntnficbk_rcd2_25_4) and downloaded ebooks (readfrq_mat_dwnld_ebk_rcd2_25_13) (excluding other forms of books asked about such as religious books, cookbooks, etc.). This variable was constructed in three frequency versions:
    a. Ever read books: readfrq_mat_bks_ever2_rcd_2_25
    b. Regularly read books: readfrq_mat_bks_reg2_rcd_2_25
    c. Frequently read books: readfrq_mat_bks_freq_rcd_2_25
- Reading initiative awareness: new variable (read_init_aware_all_rcd_3_36) generated for respondents who were aware of any of the listed reading initiatives.
- Regular community library users: new dichotomous variable lib_freq_commlib_cl_regular generated from lib_freq_commlib_cl for % of adults who are regular library users

Reader Personas

One of the innovations of the National Reading Survey 2023 is the development of Reader Personas. This section of the technical report describes the process followed to generate the Personas.

The Persona generation process included the following steps:

- Define six dimensions of reading, based on theory and the project's theory of change that reading practice is influenced by reading motivation and materials access:
    a. Reading practice: purpose,
    b. Reading practice: habits,
    c. Reading practice: volume,
    d. Reading practice: depth,
    e. Reading motivation & identity and
    f. Reading materials access.
- Generate an index for each dimension out of a set of survey variables that measure the underlying concept. This included:
    a. Selecting the variables (based on theory and the education sector experience of the data analysis team)
    b. Weighting the variables and the respective responses to each variable so that more important forms of reading carried more weight within the index (see below for weightings applied). The weighting is based on theory.

- Clustering respondents based on the six indices:
    a. All indices were generated using the weighted survey dataset, and then were normalised to carry the same weight within the clustering algorithm
    b. Missing values in any of the indices and clustering were treated in the following ways:
        i. If a respondent has data for ANY of the variables within an index, they are included in that index.
        ii. If a respondent has data for ANY of the indices, they are included in the clustering process.

Indices

The index construction recognised that not all reading offers the same benefits to the reader and to society. To build indices for each dimension, we weighted some variables more heavily than others. At a high level, the weighting within each index is described in Table 1Table 2.

Table 2: Reading Dimension Indices

1. Reading Practices: Purpose - reading for enjoyment is weighted more highly than reading for information and reading to communicate
2. Reading Practices: Habits - habitual reading (daily or several times a week) is weighted more heavily than less frequent reading
3. Reading Practices: Volume - this index combines the total time spent per week across all types of reading
4. Reading Practices: Depth - reading long text is weighted more heavily than medium and short text lengths
5. Reading Motivation and Identity - the index combines questions about the value of reading, being motivated to read and self-identifying as a reader
6. Reading Materials Access - the index combines the number, types and diversity of reading materials (print and digital), access to the internet and library use. The presence of printed reading materials in the home, and especially books, was weighted more heavily than other material types.

The index construction is described below in more detail:

Reading Purpose (why you read: enjoyment, information, communication)

- All missing values (Inability to Read) were replaced with 0 (never read)
- Variables Included:
    - 2.22.1 Read for your own personal enjoyment (like reading for entertainment or relaxation)
    - 2.22.2 Read to get information or instructions or learn something (like reading newspapers, for work, for studies, instruction manuals)
    - 2.22.3 Read and write to communicate with others (like reading letters, sms, chat messages, social media posts, emails)
- The values for these 3 variables were ordinal, hence they were recoded into numeric based on their rank. Since the intention of this index is to focus on the most prevalent type of reading (rather than the frequency of reading per se), the response options were given values as follows:
    - Never = 0
    - Less (a few times a year/less than once a year/refuse) = 1
    - About once a month = 2
    - Several times a month = 3
    - Several times a week = 4
    - Daily = 5
- Then weights were applied to each recoded variable based on their importance to the index: Reading Type Index = 3*enjoyment + 2*information + communication
- Unscaled, this index ranges from 0 - 30
- Then the Reading Purpose Index was scaled to between 0 and 1

Reading Habit (how regularly you read)

- All missing values (Inability to Read) were replaced with 0 (never read)
- There were 17 variables originally chosen to calculate the Reading habit index. However, there were 3 variables excluded as explained below.
- Variables Included:
    - 2.22.1 Read for your own personal enjoyment (like reading for entertainment or relaxation)
    - 2.22.2 Read to get information or instructions or learn something (like reading newspapers, for work, for studies, instruction manuals)
    - 2.22.3 Read and write to communicate with others (like reading letters, sms, chat messages, social media posts, emails)
    - How often do you read … (for yourself) 2.25.1 Print Magazines
    - 2.25.2 Print Comic books
    - 2.25.3 Print fiction (novels, stories)
    - 2.25.4 Print Non-fiction books (informational, documentary)
    - 2.25.6 Print Newspapers
    - 2.25.7 Magazines, blogs, opinion pieces online and on mobile apps
    - 2.25.8 Fiction (novels, stories) online and on mobile apps
    - 2.25.9 Non-fiction books online and on mobile apps
    - 2.25.11 News online and on mobile apps
    - 2.25.13 Downloaded e-books or on mobile apps
    - 3.4 How often do you read with young children?

- Variables Excluded: 3 variables were removed from the original 17 variables because a large proportion of respondents read only these types of materials daily, while not also reading in other ways. This therefore would have skewed the distribution of 'frequent' readers. The variables excluded were:
    - How often do you read … (for yourself) 2.25.5 Print Religious books
    - 2.25.10 Religious texts online and on mobile apps
    - 2.25.12 Social media posts
- The values for the 14 included variables were ordinal, hence they were recoded into numeric based on their rank. Since the intention of this index is to focus on the most prevalent frequency of reading (rather than the type of reading or the materials being read), the response options were given values as follows to disproportionately weight more frequent reading:
    - Never = 0
    - Less (a few times a year/less than once a year/refuse) = 1
    - About once a month = 2
    - Several times a month = 5
    - Several times a week = 8
    - Daily = 10
- To calculate the Reading habit index, the following logic was used:
    - Score for 2.22.1 Read for your own personal enjoyment + Score for 2.22.2 Read to get information + Score for 2.22.3 Read and write to communicate + Highest Score among remaining 11 variables for types of reading materials (if score is 10 for any of these variables, the respondent gets a score of 10 for this group of variables)
    - There is a maximum potential score of 10 for each of the four constituent parts of this index (2.22.1, 2.22.2, 2.22.3, combined score for 11 variables) so this unscaled index has a range from 0 - 40
- The Reading habit Index was scaled to 0 – 1.

Reading Volume (how much time you spend reading)

- Variables Included:
    - 2.23.1 How many hours per week do you generally spend reading for personal enjoyment (not work or study)
    - 2.23.2 How many hours per week do you generally spend reading to get information or instructions or learn something (for work, study or because you choose to)
    - 2.23.3 How many hours per week do you generally spend reading and writing to communicate with others
- Outliers were treated as described in the report section on outliers above (excluding values above 112 hours per week for each variable).
- The number of hours were grouped into buckets. The groupings were done based on an analysis of the distribution. The mean for reading for enjoyment and for information was five and the mean for reading for communication was 10 so 5-10 was chosen as 'medium'. The cut-off point between high and very high was chosen because 21 is the 90th percentile of the reading for enjoyment and reading for information distribution:
    - 0 = Never = 0
    - 1-4 hours = low = 1
    - 5-10 hours= medium = 2
    - 11-21 hours = high = 3
    - 21+ hours = very high = 4
- Then the max value (0 – 4) across these 3 variables was chosen to represent the Reading Volume Index (i.e., if the respondent reads at 'very high' volumes for any purpose, they are coded as having an overall 'very high' reading volume)
- The Reading Volume Index was scaled to 0-1.

Reading Depth (what length of texts do you read)

- All missing values (Inability to Read) were replaced with 0 (never read)
- There were 17 variables (i.e., variables 4.9.1 to 4.9.4, 2.25.1 to 2.25.13) originally considered to calculate the Reading Depth index.
- Variables Excluded: 3 variables were removed from the original 17 variables because the depth of that material type is not clear (i.e. religious texts may be long or short) or are too generic (i.e. almost all respondents read social media). The excluded variables were:
    - 2.25.5 Print Religious books
    - 2.25.10 Religious texts online and on mobile apps
    - 2.25.12 Social media posts
- Variables Included: The 14 variables included in the Reading Habit index were coded based on how frequently they were read (daily = 5, several times a week = 4, several times a month = 3, about once a month = 2, less (a few times a year/less than once a year/ refuse to answer) = 1, never = 0) and weighted based on their length/depth. Each variable therefore received a composite score of frequency x weight:

Variable	Weight
4.9.1 Short messages	1
4.9.2 Short articles	3
4.9.3 Medium text	4
4.9.4 Long text	5
2.25.1 Print Magazines	3
2.25.2 Print Comic books	3
2.25.3 Print fiction (novels, stories)	5
2.25.4 Print Non-fiction books (informational, documentary)	5
2.25.6 Print Newspapers	3
2.25.7 Magazines, blogs, opinion pieces online and on mobile apps	2
2.25.8 Fiction (novels, stories) online and on mobile apps	5
2.25.9 Non-fiction books online and on mobile apps	5
2.25.11 News online and on mobile apps	2
2.25.13 Downloaded e-books or on mobile apps	5

- To calculate the Reading Depth index, the weighted scores for the 14 variables were summed up. The max score was 255.
- The Reading Depth Index was scaled to 0 – 1.

Reading Motivation and Identity (the extent to which you value reading and consider yourself to be 'a reader')

- All missing values (Inability to Read) were replaced with 0 (never read)
- There were 15 variables (i.e. variables 2.29.8, 2.28.2 to 2.28.12, 3.25.6 & 3.25.12) originally considered to calculate the Reading Motivation index.
- Variables Excluded: 2 variables were excluded because they were only asked of respondents who live with children and so had too many missing values for an index covering all adults. The excluded variables were:
    - 3.25.12 When I was small, someone in my family read to me almost every day
    - 3.25.6 Reading together is a good way for adults (parents/grandparents/caregivers) and children to bond and build positive relationships with each other
- Variables Included: The 13 variables included in the Reading Motivation index were coded based on response:
    - self_description codes: passionate = 4, regular = 3, occasional = 2, aspiring = 1, non-reader = 0;
    - 2.28.2 to 2.2.8.12: strongly agree = 2, agree = 1, neutral = 0, disagree = -1, strongly disagree = -2. Note that negatively formulated variables had their scales inverted (see below).
    - 2.36: true = 2, false = 0

- Then the variables were weighted based on their importance in expressing reading identity (theory-based weighting). Each variable therefore received a composite score of response score x weight:

Variable	Weight
2.29.8 How would you describe yourself? Which of these categories do you identify with most: passionate reader, regular reader, occasional reader, aspiring reader, non-reader	2
2.28.2 I read only if I have to (inverted values)	1
2.28.3 Reading is one of my favourite hobbies	2
2.28.4 I like talking about books with other people	1
2.28.5 For me, reading is a waste of time (inverted values)	1
2.28.6 For me, reading is a way to explore new people and places._values	1
2.28.7 I read only to get information that I need (inverted values)	1
2.28.8 I read as a way to improve my economic situation	1
2.28.9 If people in my family saw me reading a book, they would make fun of me (inverted values)	1
2.28.10 If my friends saw me reading a book, they would make fun of me (inverted values)	1
2.28.11 Reading helps me relax	1
2.28.12 Reading is stressful for me (inverted values)	1
2.36 Which of the following people in your life ever talk about the importance of reading and storytelling?/My friends and family	1

- To calculate the Reading Motivation index, the 13 composite scores were summed up. The max score was 34.
- The Reading Motivation Index was scaled to 0 – 1.

Reading Materials Access (the amount and diversity of reading materials you engage with)

- The intent of the index is to assess the extent to which the respondent has access to a diversity of reading materials. The Reading Materials Index was calculated differently from the other indices in order to equally value different forms of materials access (e.g. print and digital) without disadvantaging those who choose not to read in one or the other.
- The Reading Materials Index has five themes, each with its own composite or 'bucket' score, which is then combined into the overall Index:
    - How many books are in your home right now, including books you own and those you have borrowed? (question 2.24.11)
        - Responses — Score
        - 0 = Never — 0
        - 1-10 books = low — 1
        - 11-20 books = medium — 2
        - 21-30 books= high — 3
        - 30+ books= very high — 4
    - Diversity of what do you have on paper in your home (calculation based on number of material types respondents mention in response to question 2.24.9). 13 variables under the theme 2.24.9 were summed which starts from "2.24.9 Books for children" to "2.24.9 Textbooks". This variable was then coded into buckets based on the distribution of the data. The buckets were as follows.
        - Responses — Score
        - 0 = None — 0
        - 1 type = low — 1
        - 2-3 types = medium — 2
        - 4-5 types = high — 3
        - 5+ types = very high — 4
    - Do you ever access free books or stories online or on mobile apps? (question 3.31) Yes = 1, No = 0
    - Whether ever read e-books (question 4.5) Yes = 1, No = 0

- Sum of responses to where respondent sources books (negative questions were inverted and yes/no questions were weighted *3 to give them equal weight as question 5.11)
    - 5.11 How often do you borrow books from the library? Frequently = 3, Regularly = 2, Rarely = 1, Never = 0
    - 3.31 Do you ever access free books or stories online or on mobile apps? Yes = 3, No = 0
    - 2.31 Where do you usually get books?/I don't want to get books No = 3, Yes = 0 (inverted scale for negative questions)
    - 2.34 If there was a popular new book you wanted to buy, where would you prefer to buy it from?/I wouldn't buy it anywhere No = 3, Yes = 0 (inverted scale for negative questions)
    - 5.5.4 Another place to borrow or use things to read for free Yes = 3, No = 0

- Each of these 5 themes were normalized individually (0-100) and then an average across all these themes was used to represent the Reading Materials Index, which was finally normalised to 0 – 1.

Clustering

The clustering algorithm used was KMeans (run in R). This was chosen because:

- There was no specific outcome variable that we were trying to predict. KMeans is used when you have a set of features you want to use to find collections of observations that share similar characteristics
- All our features were numeric. Some of our features were ordinal or binomial, but all could fit into the algorithm
- KMeans is the most popular and well-studied clustering algorithm, which allows replicability and testing by others in future.

We chose five as the optimal number of clusters because it has the smallest AIC (Akaike Information Criteria).

Regressions

The survey findings report includes the results of four regression analyses conducted. This section outlines the variables included in each of these analyses. The regressions were done in STATA and do files are available on request.

In all cases, the steps followed for the regressions were:

- Variables were reformulated as required
- Univariate analysis of all independent variables to check for collinearity using an F nested model approach, comparing the F stats and postestimation tests
- Different reference groups were used for some variables (age_group, area, household income, education level, number of books in home, etc.)
- Multiple models were tested for each regression, checking for goodness of fit.
- For the final model and reference groups, contact us at info@readingbarometersa.org to request the regression .do files in STATA.

Regression 1: What explains reading with children?

Dependent Variable: RwC_behav_rwc_at_home_3_1 (Survey question: 3.1 Which of these statements is mostly true for you... I read to children in my home)

Dependent variable type: dichotomous
Regression type: svy:logistic
Filtered by: respondents who live with children (live_with_children_1_16), N= 2071
Independent variables (entered in this order):

- Gender: gender_cl_1_3
- Number of children in the household: no_children_live_with_1_16
- Area lived in: area_lived_1_7
- Age Group: agegroup_1_1
- Highest level of education achieved: edca
- Household income (imputed): hhincome
- Number of children's books in the home: new combined variable for whether own any picture books or 'readers' (RwC_bks_num_readers_cl_3_10 and RwC_bks_num_pic_bks_cl_3_9)
- Number of books in the home (including all types of books): bks_num_all_2_24_11
- Self-identification as a reader: self_description_cl
- Whether someone in the adult's family read to them as a child: fam_read_daily_cn
- Whether the adult agrees that reading with children before they can talk helps them learn: rwc_readbeforetalk_learn_cln
- How many hours per week the adult reads themselves for enjoyment: hrspw_enjoym_rcd_2_23_1

Regression 2: What explains if adults (ever) read?

Dependent Variable: ever_read (composite variable combining if the respondent can read (inability_read_letter_cat_1_19) and if they ever spend time reading for enjoyment, information or communication (hrspw_enjoym_Cl_2_23_1==0& hrspw_info_read_Cl_2_23_2==0& hrspw_read_comm_Cl_2_23_3==0)

Dependent variable type: dichotomous
Regression type: svy:logistic
Filtered by: none. All adults included in analysis
Independent variables (entered in this order):

- Age group: agegroup_1_1
- Population group: popgroup_1
- Highest education level achieved: edca
- Monthly household income: hhincome
- Gender: gender_cl_1_3
- Area lived in: area_lived_1_7
- Reading barriers – sight: read_barriers_sight_1_21
- Reading barriers – reading disorder: read_barriers_disorder_1_21
- Reading barriers – not enough light at home: read_barriers_light_1_21
- Reading barriers – never taught to read: read_barriers_not_taught_1_21
- Self-identification as a reader: self_description_cl
- Whether live with children: live_with_children_1_16

Regression 3: What explains if adults frequently read long texts?

Dependent Variable: depth_long_text (cleaned variable based on depth_long_text_rcde_4_9_4, including missing variables skipped on inability_read_letter_cat_1_19 as 'never').

Dependent variable type: ordinal
Regression type: svy:ologit
Filtered by: none. All adults included in analysis
Independent variables (entered in this order):

- Gender: gender_cl_1_3
- Whether live with children: live_with_children_1_16
- Area lived in: area_lived_1_7
- Employment status: employment_status_rcde_1_10
- Highest education level achieved: edca
- Monthly household income: hhincome
- Age group: agegroup_1_1
- Population group: popgroup_1
- Hours per week read for enjoyment: hrspw_enjoym_rcd_2_23_1
- Hours per week use social media: hrspw_socialmed_rcd_2_23_4
- Number of books in the home (including all types of books): bks_num_all_2_24_11
- Self-identification as a reader: self_description_cl
- Would read more if friends and family talked about books: read_more_friends_fam_ct
- How frequently visit the community library: lib_freq_commlib_cl

Regression 4: What explains if adults use the library for reading?

Dependent Variable: lib_reader_composite_5_7 (composite variable combining if the respondent ever uses the community library to read books, read newspapers, read with children or borrow books: if lib_activities_Read_bks_5_7 ==1&lib_visited_ever5_1==2|lib_activities_Read_newspprs_5_7==1&lib_visited_ever5_1==2| lib_activities_readwchild_5_7 ==1&lib_visited_ever5_1==2|lib_activities_borrowbks_5_7==1&lib_visited_ever5_1==2). Respondents who ever visit the library or have not visited the library in the past year (and were therefore not asked about activities in the library) were included in the analysis but coded as 'never' having done the above activities.

Dependent variable type: dichotomous
Regression type: svy:logistic (note: Log binomial tested as alternative)
Filtered by: none. All adults included in analysis
Independent variables (entered in this order):

- Population group: popgroup_1
- Gender: gender_cl_1_3
- Whether live with children: live_with_children_1_16
- Area lived in: area_lived_1_7
- Age group: agegroup_1_1
- Highest education level achieved: edca
- Monthly household income: hhincome
- Self-identification as a reader: self_description_cl
- Would read more if friends and family talked about books: read_more_friends_fam_ct
- How frequently visit the community library: lib_freq_commlib_cl
- Number of books in the home (including all types of books): bks_num_all_2_24_11

Technical Review

The statistical approach to the regressions, indexing and clustering was reviewed by an independent statistical expert (Ling Ting) and confirmed as well-designed and well-executed.

Lessons Learned

The following lessons were learned with regards to questionnaire design:

- Variable naming in the questionnaire: in the original questionnaire, we used variable names that were not self-explanatory, leading to the need for renaming in the analysis phase. Future uses of the questionnaire should use the variable names in the clean dataset. A renamed questionnaire is provided on the National Reading Survey website.

- Skip pattern lessons
    - The main lesson regarding skip patterns was that we skipped too many other questions based on the 'ability to read' response (inability_read_letter_1_19: 1.19 Imagine that a letter arrives for you from a friend or relative. Which of the following would you do with this letter? I would ask someone to read it to me). The following questions should not have been skipped on this response:
        - Questions about ever reading for enjoyment, information or communication, to reconfirm reading practices;
        - Motivation and identity questions;
        - Questions about reading to young children. We did ask whether their older children read for themselves.
        - Questions on text length;
        - Participation in or awareness of reading initiatives. We did ask about access to (use of and awareness about) free reading materials, and some people who had previously responded that they 'can't read' responded that they are aware of such materials and some use them sometimes or frequently, which reinforces the importance of not skipping on the 'inability to read' response.
        - Similarly, we asked those who 'can't read' questions about library use and some do use the library, suggesting that other reading-related questions should have been asked of this group.

    - Some questions that were skipped on the respondent saying they did not have any books in the home (bks_num_all_2_24_11) should have been asked. Early in the survey, all respondents were asked if they had children's books or readers but not how many, and then the triangulation questions on *how many* children's books and readers they have were skipped for those who had reported not having any books. Future iterations of the survey should ask these again.

    - The question about whether the adult was read to as a child were included in the section on reading with children, and that entire section was skipped if the respondent did not have children. That question should in future be included in the section on adult reading identity and motivation and asked of all respondents.

- Language use questions
    - Questions about language *preferences* were consistently asked as multiple response options in the survey, as was the question about current language use for reading to communicate. However, the questions about actual language use for reading for enjoyment and for information were asked as single response. These should in future iterations also be asked as multiple response options.

- Reading barriers: The question about sight barriers should be reformulated to ask about "having problems with sight *and* not having access to glasses or other forms of sight improvement".

- Frequency of reading with children: The questions about frequency of reading with children should in future be asked of all adults living with children, not skipping on any previous question about 'ever' reading with children or based on age of the children. Further specification regarding age-group of the child can be added after the initial frequency questions.

- Library access: In the questionnaire, respondents were first asked if they had visited (any) library in the past year, and then asked about having access to a community library. This order should be inverted in future iterations of the survey.

- Working hours questions: if respondents said they were employed, they were asked if they worked from home or outside the house, and at what times they left home and returned home. The intention was to analyse the amount of time adults had to spend with children. These questions did not work well from a data quality perspective and so were not analysed. Future iterations of the survey may decide to exclude them or revise how they are asked.

Recommendations

The National Reading Survey is intended to promote debate and galvanise collective action by generating data that can be used to agree on sector goals and monitor progress towards those goals at a high level.

To this end, the recommendations relate to the improvement of the collective data environment.

Adding to the Data

- Literacy sector actors should pay more attention to adult reading practices and motivations, as well as asking questions about adult practices of reading with children in the home. To this end, they can use the NRS questionnaire, or shortened versions of it, as part of exploratory studies or as part of monitoring and evaluation studies (baseline and endline surveys).
    - The questionnaire is freely available under a Creative Commons BY-NC-SA license and can be adapted as required, with acknowledgement of the original designers. We request that organisations that use the questionnaire:
        - Retain the variable names and question formulations as much as possible, so that findings can be compared, although the number of questions can be shortened, the order changed as needed and additional questions added as required for the specific study.
        - Let the NRB team know that the questionnaire is being used and how it has been adapted: info@readingbarometersa.org
        - Share an anonymised version of the dataset with the NRB team so that a composite picture of adult reading practices in different parts of the country and in different contexts can be compiled between iterations of the nationally representative NRS.

Using the Data

- The NRS 2023 dataset if publicly available for further analysis and can be downloaded from DataFirst: South Africa - National Reading Survey 2022-2023 (uct.ac.za). We strongly encourage further analysis of the dataset. Some suggested areas for further analysis include:
    - Relationship between economic status and materials access: Deepen analysis of reading materials access (materials in home, books in home, book sources, etc.) in relation to household income and employment patterns. Statistically explore the extent to which the impact of income on reading practices is mediated via materials access.
    - Libraries: Combine NRS respondent geolocation data (not included in public dataset but can be requested from NRB team) with Libraries geolocation data to investigate the relationship between respondent distance from a community library and their level of awareness of/reported access to a library. Consider recommendations for library placement based on this analysis.
    - Teen reading: Compare teen self-reported reading practices (16-18 year old survey respondents) to caregiver reports about teen reading practices (15-18 year old children living in homes of respondents).
    - Religious reading: further explore how religious reading fits into broader reading cultures by looking at the reading identities, socio-economic backgrounds and other reading practices of people with high levels of religious reading.

Learn more

For more information about the National Reading Survey and the National Reading Barometer, visit www.readingbarometersa.org to access:

- The summary and full survey findings reports
- The full dataset and questionnaire (also downloadable at DataFirst: South Africa - National Reading Survey 2022-2023 (uct.ac.za)
- "What kind of a reader are you?" quiz
- Special issue briefs on reading with children, digital reading, languages and libraries
- Latest news and analysis

Follow the National Reading Barometer project on social media for the latest updates:

- LinkedIn: @national-reading-barometer-south-africa
- Facebook: @ReadingBarometerSA
- Twitter: @BarometerSA
- Instagram: @readingbarometersa
- YouTube: @Readingbarometersa